DOCUMENT RESUME

ED 212 087 EA 014 365

AUTHOR White, Karl; And Others
TITLE State Refinements to the ESEA Title I Evaluation and Reporting System: Utah 1979-80 Project. Final Report.
INSTITUTION Utah State Office of Education, Salt Lake City.
SPONS AGENCY Department of Education, Washington, D.C.
PUB DATE May 81
CONTRACT UTP-301-1
NOTE 236p.

EDRS PRICE* MF01/PC10 Plus Postage.
DESCRIPTORS *Achievement Tests; Elementary Education; Evaluation Methods; Norm Referenced Tests; Questionnaires; Scores; State Surveys; Tables (Data); Teacher Attitudes; *Test Format; *Testing; Test Theory; *Test Validity; *Test Wiseness
IDENTIFIERS Elementary Secondary Education Act Title I; *Title I Evaluation and Reporting System; *Utah

ABSTRACT

To explain discrepancies in Utah's elementary school test results under the Elementary and Secondary Education Act's Title I Evaluation and Reporting System (TIERS), researchers investigated the adequacy and validity of TIERS evaluation models. Model A (norm-referenced testing) is used in most Utah school districts, in preference to Models B or C (both involving comparisons with control groups). The researchers reviewed previous research and conducted four projects that (1) compared test results under Models A and B, (2) assessed how well Utah's tests met Model A's assumptions, (3) analyzed the effects of test formats, and (4) examined the impact of training students and teachers in test taking and administering. Using tests, interviews, and observation, the projects analyzed test scores, educators' attitudes, and students' and teachers' test behaviors in several school districts, especially the Salt Lake City School District. The results indicate that (1) Model A overestimates Title I's impact, (2) most of Model A's assumptions are met, (3) test format heavily affects test results, and (4) training students and teachers to take and give tests improves scores. Ten appendices reproduce cover letters and data collection forms. (RW)

FINAL REPORT

STATE REFINEMENTS TO THE ESEA, TITLE I EVALUATION AND REPORTING SYSTEM: UTAH 1979-80 PROJECT

MAY 1981

Submitted by
UTAH STATE OFFICE OF EDUCATION
Salt Lake City, Utah

FINAL REPORT

State Refinements to the ESEA Title I Evaluation and Reporting System: Utah 1979-80 Project[1]

by

Karl White
Cie Taylor
Larry Carcel
Nancy L. Eld
Utah State Unive

in cooperation with

Utah State Office of Education

May, 1981

[1]This work was supported by the United States Department of Education under Contract #UTP-301-1. Opinions included herein do not necessarily reflect official Education Department policy.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

CHAPTER I

OVERVIEW OF THE STUDY

During the 1979-80 school year, Title I services were provided to almost 20,000 students in approximately 240 schools throughout the 40 school districts in Utah. The impact of the Title I programs on students' achievement in these districts was measured in accordance with the Title I Evaluation and Reporting System (TIERS), which had been developed under contract to the U.S. Department of Education.

In implementing TIERS, each district in Utah chose one of three models (Model A - norm referenced, Model B - comparison group, or Model C - special regression)[1] to evaluate the impact on student achievement of Title I programs. Each of these models compares the achievement level of students at the conclusion of the Title I project with an estimate of the achievement level which would have resulted if the students had not participated in Title I. According to the developers of the evaluation models which underlay TIERS, each of the three models should yield comparable results if properly implemented (Tallmadge and Wood, 1976). Such comparability is essential if data from different programs are to be aggregated to determine the statewide or nationwide effect which Title I programs are having on student achievement.

Districts in Utah have been using TIERS to some degree as a part of their Title I evaluations since the 1977-78 school year. During that year, approximately 50% of the state's 40 districts used the USOE models to examine student achievement resulting from Title I programs. In 1978-79, 90% of the

---

[1]These three models are not described in detail since they are very familiar to most readers. Additional explanation of the models can be found in Tallmadge and Wood (1976).

districts used one of the models, and during the 1979-80 school year, all of the districts in Utah used one of the evaluation models proposed by TIERS, although only 33% of the districts were required to report their results to the state. As is the case nationwide, most districts in Utah have selected to implement Model A (the norm referenced model).[2] During the 1979-80 school year, Model A was used in 36 of the 40 Utah school districts.

Problem Statement

An analysis of the initial results from Title I evaluations in Utah, which have used Model A, has raised a number of questions with Utah State Office of Education (SEA) personnel who are responsible for the statewide implementation, coordination, and evaluation of Title I within Utah.

First, during the past five to six years, SEA Title I personnel have systematically worked to identify which Title I projects are most effective. The identification of effective programs has depended on data from standardized achievement testing, on-site visits to the projects, and interaction with project staff. The experience of SEA personnel in conjunction with standardized test data have provided a fairly good indicator of which districts have historically operated the "best" Title I programs. The results from the Model A evaluations, however, sometimes contradict these historically based assessments of program quality. Some of those programs which have traditionally been the "best" programs have shown very limited or no gains, and some of those programs which have always been considered by SEA

---

[2] Model A can be implemented using either a nationally normed standardized test (Model A1) or a Criterion Referenced Test (Model A2). Throughout the remainder of this paper, "Model A" refers to Model A1.

personnel to be the weakest programs have demonstrated the best gains. In many cases, this seems to be occurring even though the individual programs and associated personnel have not changed appreciably.

Secondly, unless there are dramatic changes in either the type of student being served or in the nature of Title I programs being provided to children, it would seem that the impact of Title I within a given district should not shift dramatically from year to year. However, in some districts where no such changes have occurred, the impact of Title I programs since Model A was implemented seems to have changed substantially and to have shifted back and forth from year to year.

In view of the fact that most districts in Utah have chosen to use Model A because of its greater feasibility, these initial results from the model caused SEA Title I personnel, as well as many local district personnel, to become concerned about the validity of Title I evaluation results obtained using Model A. This project was undertaken to provide further information and explanation of the apparent discrepancies of Title I evaluation data prior to and after the implementation of TIERS. The intent of such a project was to allow both SEA and LEA Title I personnel to proceed with more confidence in making decisions regarding the impact of Title I projects.

In considering the reasons for the inconsistencies in evaluation results a number of factors were identified which might be partially responsible for the apparent discrepancies observed with the Model A evaluation model. Each of the factors is discussed briefly below.

Inaccuracy of reported information. It is possible that local schools and districts are not reporting data accurately. The TIERS models have numerous areas in which arithmetic, procedural and/or clerical errors could occur which would substantially alter the results for a given project. Stonehill and English (1979) reported that in studies conducted by the

Department of Education and various Regional Technical Assistance Centers (TACs), more than 95% of the districts made some errors when producing their LEA Title I evaluation report.

This source of error (i.e., the accuracy of reported data) has been a source of concern to SEA personnel in Utah since districts began to implement the evaluation models. Efforts to reduce errors of this type constitute a major portion of the responsibilities of Utah SEA Title I personnel. For example, much of the hand calculation by LEA personnel has been eliminated through the development of a computer program at the SEA which uses raw data submitted by LEAs to do most of the necessary computation. Workshops and various standardized forms for reporting raw data have also been used to further reduce errors. In addition, much of the assistance provided to Utah by the Region VIII TAC has been targeted on this area. As a result of these past and ongoing activities, there was reasonable confidence among SEA Title I personnel that arithmetic, procedural and/or coding errors would not be a major source of error in 1979-80 Title I evaluation data in Utah.

Violation of Model A assumptions. A second potential source of error in the results of Title I evaluation data obtained using Model A was that some of the assumptions made to ensure "proper implementation" of the model were being violated. Briefly summarized, the assumptions made by Model A include:

1. Publisher's directions for administering and scoring tests are adhered to closely.
2. Selection tests are separate from pretests in order to eliminate regression towards the mean.
3. Both pre- and posttests are administered close to empirical norm dates and appropriate adjustments are made where necessary.

4. An appropriate level of the test is used in order to avoid floor and/or ceiling effects.
5. Make-up tests are given within two weeks to students who were absent on testing day.
6. The norms of the tests being used for pre- and posttests are appropriate for the LEA.
7. Test content in the standardized tests appropriately matches the instructional emphasis of the LEA.

Violations of one or more of these assumptions by users of Model A may bias results either positively or negatively. The biasing effect of such violations has been discussed to some degree by other writers (Conklin, 1979; Kaskowitz and Norwood, 1977; Linn, 1978; Long, Schaffran, and Kellogg, 1977; Murray, 1979; Tallmadge and Horst, 1977). One purpose of this study was to explore the frequency and estimate the probable consequences of such violations in Utah school districts which are using Model A.

Validity of results when Model A is properly implemented. A third possible explanation for the observed discrepancies in the TIERS Model A is that Model A may not provide an accurate measure of program impact. The theoretical adequacy of Model A has been questioned by a number of writers (Kaskowitz and Norwood, 1979; Linn, 1978) with particular concern being expressed about the equipercentile assumption--i.e., the assumption that if the Title I program is completely ineffective, students enrolled in the program will maintain their same percentile ranking with respect to the norm group from pretest to posttest.

With a major effort already being directed towards solving the problems created by inaccurately reported data, an examination of the frequency and probable consequences of violations of model assumptions, and an examination

of the validity of Model A was intended to provide SEA and LEA personnel in Utah and other states with useful information for selecting an appropriate model and interpreting the results of Title I evaluations.

Objectives

The overall goal of this study was to investigate and draw conclusions about the adequacy and validity of Model A as it is currently being implemented in Utah school districts. This goal was addressed through the accomplishment of the following objectives.

OBJECTIVE #1: Compare the Estimates of Achievement Growth Due to a Title I Project Using Model A and Model B with the Same Group of Title I Students (Grades 2-4)

Since Model B is generally regarded as the most rigorous of the three models, a properly implemented Model B should provide the best indicator of the true impact of a Title I program on students' achievement. A comparison of the results from the two models would yield valuable information about the validity of Model A results.

Previously completed studies had compared the results of two different models using actual data (Faddis and Arter, 1979; Davidoff, 1978; Kearns, 1978; Storlie, Rice, Harvey and Crane, 1979; Stennan and Raffield, 1977) or simulated data (Echtenacht, 1978; Mandeville, 1978; Stennan and Raffield, 1977). Most of these studies have not included Model B as one of the models being compared. Faddis and Arter (1979) compared results of Model B and Model A and found that Model A resulted in lower estimates of growth than Model B. Although their results were related to the objectives of this study, Faddis and Arter (1979) only considered ninth grade students, did not explicate adequately how the control schools for Model B were selected, and had significant dropout rates (55%) between pre- and posttests.

OBJECTIVE #2: Analyze the Title I Evaluation Procedures and Results from Utah Districts which have Used Model A to Pinpoint the Frequency and Probable Consequences of Violating Model A Assumptions

This objective was addressed to provide empirical evidence about the frequency and extent to which the assumptions made by Model A were being violated in Utah school districts. Data about violations of Model A assumptions was intended to provide information for estimating whether the inconsistencies regarding Model A evaluation results could be due to improper implementation of the model. The assumptions considered in this analysis were referred to earlier.

Additional Studies

As the work scope defined in the contract with the Department of Education commenced, a number of other activities were identified which related to objectives of the original work scope. A number of these activities were undertaken and accomplished by project staff. Some of this work was partially funded by another small grant from the Office of Special Education (a student-initiated research grant conducted by Cie Taylor and Karl White), other parts were paid for by research funds from Utah State University, and some parts were made possible by more efficient use of project resources. This efficient use of resources was possible because project staff were already working with staff members in a number of schools in the Salt Lake District and many of the same procedures and instrumentation developed for this project could be used in the additional studies. In other instances, some of the additional work was initiated because once the project had begun, members of the research team had new insights about what kinds of information would meaningfully impact on the problems which had originally motivated this study.

Consequently, the research reported in this final report is much broader than the workscope originally outlined in the proposal. All of the originally described activities were accomplished and the report will summarize the activities and accomplishments. In addition, the report describes the outcomes and benefits accomplished as a result of the additional projects which were made possible because of the existence of the State Refinements contract.

These additional projects are discussed in Chapter V (The Effect of Item Format on Student's Standardized Reading Achievement Test Scores); and Chapter VI (The Effect of Standardized Test Performance of Training Students; Training Teachers, and Motivating Students). The following three chapters consist of: Chapter II: A Review of Research on Assumptions Underlying TIERS; Chapter III: A Comparison of Models A and B; Chapter IV: Degree to Which Assumptions Made by Model A are Met in Utah Title I Evaluation. Each of these chapters includes procedures, results, and conclusions. The final chapter of this report synthesizes conclusions from each of the preceding five chapters into a brief summary and recommendations concerning the implementation and operation of TIERS in Utah Districts.

CHAPTER II

REVIEW OF RESEARCH ON ASSUMPTIONS UNDERLYING TIERS MODELS

Title I of the Elementary and Secondary Education Act (ESEA) of 1965 was a massive action on the part of the Federal government to upgrade the educational system of the United States. Basically, funds are provided for educational programs aimed at improving basic educational skills of educationally disadvantaged children. These skills are defined as competencies in reading, math, and language arts. Financial aid is allocated to counties with high concentrations of poor families. However, Title I students are selected upon the basis of educational need (Stonehill & English, 1979).

Since its inception, ESEA has mandated that schools receiving Title I funds evaluate their programs on a yearly basis. Throughout the early years of Title I, evaluations of a kind were performed by those receiving funds. However, by the early 1970s it was apparent that compilation of these data into a summary of educational achievement on the part of Title I students was impossible. In short, there was no way to measure the impact of a program, which over an 8-year period had cost over $10 billion (Barnes & Ginsberg, 1978).

In 1974 Congress passed Section 151 (now Section 183) of Title I which required the development of standards for evaluation which would yield comparable data across programs. Technical Assistance Centers (TAC) were also established to assist State and Local Educational Agencies in the implementation of such evaluation (Stonehill & English, 1979). Shortly thereafter the Office of Education contracted with RMC Research Corporation under competitive bidding procedures to develop the Title I Evaluation and Reporting System (TIERS).

RMC developed three basic models of evaluation. Each includes pre- and posttesting, a method for generating gain estimates in the absence of a Title I program, and procedures for converting test results into "NORMAL CURVE EQUIVALENTS" (NCEs) to permit aggregation of data across projects. These models are designed to answer the question, "How much more did pupils learn by participating in the Title I project than they would have learned without it?" (Tallmadge & Wood, 1976). However, since TIERS' inception, increasing numbers of evaluators have questioned whether or not the system works.

From the beginning TIERS has elicited heavy criticism as well as support. There is no consensus on its use, applicability, or validity. There appear to be three types of concerns with the system. The first is whether the statistical assumptions of each model are valid. The second is whether proper implementation of a given model can occur at the local level. The third is whether the models are comparable. Currently, there is no integration or analysis of the literature pertaining to these concerns. Since Title I programs cost billions of tax dollars and affect millions of children, valid evaluation is essential. This critical review will examine the TIERS models by reviewing the literature which assesses TIERS. The results of the review should be useful in determining whether the system is valid as an evaluation system.

MODEL DESCRIPTIONS

All three evaluation models use the same definition of treatment effect. Specifically, the project's impact is the actual post treatment performance minus the expected no-treatment performance. The models differ in how the no-treatment expectation is generated, but all models use pre- and posttesting to determine treatment effect.

Model A (the norm referenced model) uses test publisher's norms to generate the no-treatment expectation on the assumption that the group's ranking on the pretest would be maintained on the posttest, if there was no treatment. Model B (the control group model) uses a control group which theoretically receives an identical education as the treatment group except that they are not Title I project participants. In Model C (the special regression model), the no-treatment effect is derived by finding the treatment group's mean pretest score on the comparison group's post on pretest regression line. Each model can be used with either normed (type 1) or non-normed (type 2) tests. When non-normed tests are used, a normed test must also be administered in order to convert measured gain into NCE units.

The following sections overview each model and the associated procedures.

Model A. When nationally normed tests are used, the no-treatment expectation is the percentile status of the treatment group at the time of the pretest. This model assumes that if there is a gain due to treatment, the group will rise in percentile status. If there is no treatment effect, the percentile status will remain the same. The observed posttreatment performance is the percentile status of the group's mean posttest score, while the expected no-treatment performance is the pretest percentile.

When non-normed tests are used, a nationally normed test must be given at pretest time. Score equivalencies between non-normed pretest and normed test are determined. The median scores on the non-normed posttest are converted to normed test counterparts on the basis of these data and this figure is converted to a percentile and becomes the observed posttreatment performance. Again, the pretest percentile is the expected no-treatment performance.

Since all calculations are performed on the basis of the publisher's norms, several important factors must be observed. First, all testing (pre and post) must be done on dates within two weeks of the empirical norming date for the test or within six weeks, if publisher's norms are adjusted. Secondly, the group's composition must be similar to that of the population used in the norming sample. Third, testing must be performed exactly as done for the norming sample. Fourth, to avoid statistical regression towards the mean, the pretest may not be used to select students for the program.

Model A is the most frequently used model. It is the easiest and cheapest to implement. In both Model B and C, the calculations are more complex and the cost is much higher, since control groups must also be tested.

Model B. With this model, a control group is formed. The control group's posttest percentile (Model B1) or mean raw score (Model B2) is the expected no-treatment performance whereas the treatment group's posttest percentile or mean raw score is the observed posttreatment performance.

The pretest scores verify the groups' equivalency. If these scores differ between groups, two statistical techniques can be used to adjust for this inequality. Where random assignment of a population into groups is used, analysis of covariance is used for adjustments. When the groups are more appropriately regarded as samples from two different populations, the principal axis method of adjustment is used.

Again, several factors must be observed if the model is to be properly implemented. First, both groups must be tested at the same time, in the same manner, with the same test and level of test. Secondly, group composition must be similar in terms of socioeconomic status (SES), sex, and race. Even though small systematic differences can be adjusted for, large differences may

invalidate the model. Third, with the exception of the Title I treatment, the educational experiences of both groups must be the same.

Model B is difficult to implement. Appropriate control groups are usually not available, since it is usually impossible to randomly assign children to treatment and control groups. Moreover, very few LEAs have large groups of eligible children who are not receiving Title I services. Model C takes this factor into account.

Model C. With this model, participation in the program is based on a strict cutoff score on the pretest. All those above the cutoff form the comparison group and all those at or below form the treatment group. Post-on-pretest regression lines are calculated separately for each group. The treatment group's line represents the observed mean posttest performance corresponding to various pretest scores, i.e., the observed posttreatment performance. By projecting the comparison group's line below the pretest cutoff score, the expected no-treatment performance is obtained. The actual treatment effect is the difference between the lines measured at two points: the treatment group's mean pretest score and the cutoff score.

This model requires that the pretest-posttest relationship is linear. The two measures must be highly correlated, and no floor or ceiling test effects may be present which could create curvilinearity. Furthermore, the pretest must be used as the sole basis of selection.

REVIEW OF THE LITERATURE PERTAINING TO MODEL ASSUMPTIONS

For each model, a set of assumptions exist. These assumptions can be classified into two categories: those held in common by all three models and those specific to a given model. Although a large body of literature has

examined the assumptions made by TIERS and discussed the various problems and biases introduced upon violation of a given assumption, this literature has not been effectively summarized. The following section will detail the assumptions and examine the pertinent literature regarding assumption violations.

A. ASSUMPTIONS COMMON TO ALL MODELS

Quality control regarding test administration and data reporting is probably the single largest assumption affecting implementation of TIERS. Without quality control, all of the other assumptions are irrelevant. Quality control is defined as the accurate collection and reporting of all data necessary to implement any TIERS model. Therefore, it enters into the system at numerous places and can take many forms. The following discusses the various places where quality control affects the system.

Selection of Tests

Several decisions must be made when a test is being selected to evaluate a Title I program. First, does the test measure the curriculum being taught? Secondly, should a normed or non-normed test be used? Third, which level of the test should be administered? Finally, when should evaluation of the program occur?

Curriculum-Test Matching

On the surface, it would appear as though the objectives of Title I--to teach basic reading and math skills to educationally disadvantaged children--would allow for the use of the same instrumentation nationally. This would permit exact comparability of results, the area where many of

the problems of TIERS are encountered. However, as Tallmadge and Horst (1978) have pointed out, the basic skills often require several years to acquire (e.g., reading comprehension), and an evaluation system must either allow for the complete acquisition or confine evaluation to what is being taught during the specific period under evaluation.

Instructional programs tend to be focused on some subset of the basic skill. Thus, if testing occurred over the entire collection of subsets, gains one very specific area might not be detected. For example, suppose District A's reading program focuses on vocabulary and uses a test which emphasizes word attack skills--gains in vocabulary might go undetected. However, if a different achievement test were used (one which emphasizes vocabulary), gains might be observed. Using a test specific to the curriculum, appropriate evaluation can occur; since the more closely a test corresponds with the skill being taught, the greater the likelihood that student gains can be detected (Fagan & Horst, 1978).

While fitting a test to a curriculum may sound easy, specific analysis of, for example, a reading achievement test into its component subsets is seldom done. The importance of such an analysis was demonstrated by Porter, Schmidt, Floden, and Freeman (1978) who classified the items of the mathematics subtests of four standardized achievement tests (SAT, ITBS, MAT, and CTBS) into several factors including the nature of the material. Their analysis revealed major differences between the tests.

The major objection against the use of the same instrument nationally is that it would not allow Local Education Agencies (LEAs) or State Education Agencies (SEAs) sufficient control over their curricula. LEAs would be tempted to design curriculum to fit into the evaluation system rather than teaching to the students' needs. If the LEA focused

instruction for some students in areas not included on the national test, gains in these areas would go unmeasured. Furthermore, for the vast majority of educators, anything which would reinforce the concept of a nationally mandated education curriculum is distasteful and threatens educational and curriculum progression.

The aspect of curriculum-test match is vital to all the models. If test selection is done without considering the curriculum it is to evaluate, serious threats to internal validity occur. Tests must not be chosen because the LEA possesses a copy, or because it is inexpensive, or because the administrator likes a particular test, but because the test fits the curriculum.

Normed vs. Non-normed Tests

The second problem of test selection is deciding what type of test to use: normed or non-normed. All models have provisions for the use of non-normed tests provided that a normed test is given either at the time of the pretest (Model A2) or at some time during the period of evaluation (Models B2 and C2).

The use of a non-normed test creates several problems for the evaluator. First, there are additional costs since a normed test must also be given to establish gain estimates in terms of NCEs. These costs would be least in Model A, since only the Title I population is tested, and greater in Models B and C, where both treatment and control comparison groups also have to be tested. The second problem is one of within model comparability, e.g., will Model A, implemented with normed tests, (Model A1), yield the same estimates of gain as if implemented with non-normed tests (Model A2)? Third, non-normed tests may be overly restrictive and measure a very small subset of skills which do not reflect actual gains

made in the total area of basic reading, mathematics and/or language arts skills. However, some authors (Tallmadge & Horst, 1977; Arter & Estes, 1978) suggest that the benefits of using a non-normed test may outweigh the difficulties. These benefits include greater curricula sensitivity and greater gain sensitivity.

A non-normed test is usually a criterion-referenced test (CRT). A criterion referenced test is particularly appropriate when gains are to be measured in isolated skill areas, where a norm-referenced test (NRT) may be insensitive to the specificity of the curricula. Additionally, norm-referenced tests may be insensitive to the small gains made in a given project even though these gains could be educationally important (Tallmadge & Horst, 1977). Finally, in the case of Model A, local norms may differ from national norms. If the local population is below average, a year's worth of gain for that population might be less than a year's gain in the norming population. Those classified as educationally disadvantaged within the local population would be at even greater disadvantage when their gains were compared to the national norms (Arter & Estes, 1978).

One aspect of the above was investigated by Long; Horwitz, & DeVito (1978). While the study was specific to Model B, their conclusions are applicable to all models. National group standard deviations were estimated in a local criterion-referenced test using the local group's standard deviations on the normed and non-normed tests. This assumed that the ratio of local to national standard deviations was identical for both tests. Their results demonstrate a higher estimated standard deviation for the national norm group than the actual national norm group reported by the test publisher. The authors noted that it is uncertain whether the

error of estimation is systematic or random, and that until this issue is clarified, the use of non-normed tests to evaluate Title I projects is questionable. They suggest a system for establishing confidence intervals around the estimated standard deviation.

Comparability Within Models

Fishbein (1978) pointed out that if the CRT scores are not normally distributed, the linear transformation of these scores into NCEs will not result in a normal distribution. Since the resulting NCE distribution will be a non-normal curve, and the NCE distribution for the normed test is a normal curve, the conclusions based on these two distributions may be different. Even if the same general model (e.g., Model B) was applied to the same population tested on normed (B1) and non-normed (B2) tests, different results might be obtained. Studies by Fish (1979) and Storlie, Rice, Harvey, and Crane (1979) have attempted to make direct within model comparisons.

Fish compared Model A1 to A2 and found that both yielded similar estimates of gain as long as the two tests measured comparable skills. Unfortunately, Fish did not include enough information to critically evaluate the methodology of the study or the comparability of the tests. The User's Guide specifically states that unless a high correlation ($r \geq .6$) exists, it is inappropriate to use the non-normed test, since NCEs are yielded by extrapolating through the normed test. This correlation was a problem in the study by Storlie et al. (1979) which attempted to compare A1 and A2. In this case, the authors decided that a correlation of .56 was too low to justify equating the tests. The authors recommend a simulation study due to difficulties encountered using empirical data. This raises an additional problem with using a non-normed test. The

User's Guide presents no information for how to salvage an evaluation which, after all of the data have been collected, has a correlation lower than .6 between the NRT and CRT. This means that it is not implausible that an LEA would commence their TIERS and discover too late to change directions that the correlation between the normed and unnormed test was too low.

Content tested by CRTs. The final problem, the narrowness of content tested by most CRTs, is discussed by Fishbein (1978). This may also be viewed as a curriculum-test match problem. If objectives of a program are narrowly defined and large gains are demonstrated using CRT instrumentation, one would conclude that the Title I pupils had shown more growth than they would have without Title I. However, in terms of treatment effect, it may be difficult to determine whether or not the overall objectives of Title I were met. For example, even though the evaluation may demonstrate that large gains have been made in phonics, the students may not be better in general reading skills than they would have been without the program--phonics skills may have been improved at the expense of more general reading skills (Tallmadge & Horst, 1977). The problems created by the narrowness of CRT content can be reduced to a problem of policy vs. program objectives (Fishbein, 1978); where, a policy objective (e.g., the production of better readers) is more difficult to evaluate than a program objective (e.g., the acquisition of phonics skills).

In summary, the use of non-normed tests appears to be unsupported by the literature. Despite attempts by several authors (Tallmadge & Horst, 1977; Arter & Estes, 1978) to point out existing benefits, previous research has not demonstrated that the benefits outweigh the difficulties. The major difficulties, in addition to increased costs, lie in three areas: lack of within model comparability, lack of a normal distribution of CRT scores, and the narrowness of curriculum tested by most CRTs.

Out-of-Level Testing

The question concerning which level of a given test is most appropriate has generated some discussion in the literature (e.g., Roberts, 1978; Ozenne, 1978; Johnson & Thomas, 1979). The correct level of a test is one on which the fewest children score at either chance level or at the top score (Johnson & Thomas, 1979). Testing "out of level", in the case of Title I projects, means testing at the functional level of the child. Testing a Title I child enrolled in the fifth grade at her functional level might mean using the third grade level of the test.

Use of the incorrect level of a test can affect evaluations in two ways. If a preponderance of children score at chance level on the pretest, the pretest average is artificially inflated since the children would have scored lower had there not been a floor effect. Upon posttesting, the observed gain would be smaller than was actually the case. If a large number of students were to score at the top of the posttest, a similar effect is seen: the visible gain is less than the actual gain since the student's posttest level is actually higher than the test was able to measure. Since the test publisher's "recommended" level is not necessarily appropriate for students with very low performance levels, as would be the case in Title I classes, out-of- level testing (i.e., the use of a test level other than specifically recommended by the publisher) is often employed (Johnson & Thomas, 1979).

Out-of-level testing most seriously affects Model A, for which there is no control group, but Models B and C can also be affected. For example, if the treatment group in Model B "tops out" on the posttest, comparison to a control group in which there is no ceiling effect will not show true gains. With Model C, a change in level between pre- and

posttesting could affect the regression lines, since even minor floor or ceiling effects on either test will result in curvilinear regression lines making interpretation difficult (Estes & Anderson, 1978; this aspect will be discussed more fully under Model C assumptions).

RMC Research Corporation realized the difficulties caused by floor and/or ceiling effects in out-of-level-testing. Roberts (1978) suggested a formula for estimating the occurrence of floor and ceiling effects. Specific suggestions were given for predicting when these effects would occur as well as detecting the presence of the effects from score distributions. For example, if student score distributions on a given pretest predict that ceiling effects will occur on the posttest, Roberts suggested that a higher level posttest be given, even though (1) this violates a recommendation in the use of the same level pre- and posttest for Model A, and (2) there may be content changes between the levels.

Ironically, Roberts' (1978) recommendations may make the implementation and interpretation of TIERS more rather than less difficult. Ozenne (1978) demonstrated that two levels of the CAT, both recommended for fourth graders, yielded different results (31st percentile vs 25th percentile on the average) when given to a group of fourth graders split into random halves. Depending on the test level given, different results would be obtained and different placement might occur. Additionally, while a student's status on one level was fairly constant across time, status changed if tested on the other level. Ozenne (1978) recommended that evaluators avoid changing test levels from pre- to posttest. If change is unavoidable, Ozenne (1978) recommended double testing at posttest to adjust for any disparity between the two test levels.

Several authors have attempted to compare students' gains by using both at-level and out-of-level tests (Slaughter & Gallas, 1978; Ozenne, 1978; Long, Schaffran, & Kellogg, 1977; Crowder & Gallas, 1978; Powers & Gallas, 1978). The results of these studies are equivocal, not only across studies but within studies.

Long, Schaffran, and Kellogg (1977) attempted to determine whether elementary school Title I students would receive the same grade equivalent (G.E.) scores when tested in level and out of level. The results showed that at the second and third grade levels, in-level tests resulted in lower G.E. scores and more students eligible for Title I than if out-of-level tests were used. At the fourth grade level, the opposite was true. For example, on vocabulary, 77% of second graders and 88% of third graders would have been eligible for Title I on the basis of in-level testing, versus 54% and 66% (respectively) if selected by out-of-level testing. For students in fourth grade, however, eligibility on the basis of in-level testing was 77%, while 89% was eligible based on out-of-level testing. For all grades, out-of-level testing demonstrated greater gains than in-level testing.

Murray, Arter, and Faddis (1979), in commenting on Long et al. (1977), stated that comparable results would only be valid if both levels were appropriate, which was not the case. Additionally, they stated that the probable reason for differing G.E. scores was that students tended to score at chance level in in-level tests, but not on out-of-level tests which were suited to their functional level. While this may account for the results of grades 2 and 3, it does not account for the results of grade 4. Perhaps, a better explanation lies in the material being tested and its comparability to the students' curricula. For example, if second

and third grade Title I students were being taught at functional level, they would have performed better on the out-of-level test, while if the fourth grade Title I students' curriculum was matched better with the at-level curriculum, they would have performed better on the at-level test.

The problem of curriculum-test matching is also apparent in the Crowder and Gallas (1978) study which investigated whether standard scale scores are comparable for in-level and out-of-level tests, and whether the floor effect of in-level testing was evident in out-of-level testing. The authors suggested that for students scoring at floor level on a higher test level, the lower level test would give lower average standard scores; for students in the functional range of each test, the standard scores would be the same; and for students at the ceiling on the higher test, standard scores would be at the ceiling on the lower test. However, these patterns are not evident in their results. While the authors explained that these results were due to the relatively large increase in standard scores at the extremes of the scale transformation, Murray et al. (1979) suggested that the differences were due to differences in curriculum between the test levels used.

Powers and Gallas (1978) examined students in fourth, seventh, and ninth grades tested in level and out of level. While fourth graders attained higher percentile rankings on the in-level test (in-level 27%, out-of-level 16%), seventh and ninth graders attained higher rankings on the out-of-level test (in-level 16th and 23rd percentiles respectively, out-of-level 13th and 16th percentiles respectively). However, no significant differences were found in expanded standard scores. It can be concluded that out-of-level tests may be more precise, even though both tests provide similar information on student status.

One particular problem which has not been addressed in the literature is the use of out-of-level tests with Model A, the norm referenced model. Since publishers' norms on a given test are for a particular population (e.g., third graders), using that test on a different population (e.g., fifth graders) creates difficulties in norm referencing. If the in-level (e.g., fifth grade) test was used, a floor effect might be seen, so the evaluator is caught in a double bind. Yet, nearly all of the existing literature recommends testing at functional level, i.e., out-of-level testing. Certainly, this aspect of out-of-level testing requires further study.

In summary, there does not appear to be a strong data base supporting out-of-level testing for Model A, although much of the literature condones its use. The suggestion on the part of RMC Research Corporation that CRTs be used with Model A is particularly alarming since norms are applied to a different population than that of the norming sample. More research must be done in this area. With Models B and C, it is clear that major problems can occur if the test level is changed from pre- to posttest, as the content levels may differ. If a different level of test is unavoidable, both levels should be given at the posttest in order to adjust for disparity between the levels. Obviously, since Models B and C have local control/comparison groups, the problem of out-of-level testing is not as great as with Model A, provided that both treatment and control groups receive the same level of the test and that the test level does not change from pre- to posttest. However, it is critical to avoid floor and ceiling effects in either group.

Time of Testing

The last issue of test selection is the date of testing. One aspect of this will be discussed under the assumptions of Model A, which requires that testing be done on the same dates as the publisher's norms. The other aspect, that of the length of the evaluation period, applies to all models. The question is whether pre- and posttesting should be done on a yearly basis (i.e., test once a year so that last year's posttest is this year's pretest), or on the basis of the academic year (i.e., pretest in the fall, posttest in the spring). An important question is whether loss or gain occurs over the summer vacation and whether Title I students lose more or gain less over the summer than non-Title I students.

One difficulty in comparing spring-spring vs. fall-spring testing is that most publishers' norms tend to be based on spring-spring testing (Conklin, 1979). This means that to obtain norms for fall performance, empirical norms must be interpolated (unless empirical norms exist for both fall and spring, which is the case for some tests). Conklin (1979) and DeVito and Long (1977) criticize such interpolation since it usually involves the equating of different test forms or levels and assumes that different grade levels have the same pattern and rate of growth.

The User's Guide does not give instructions as to how interpoltation is to be performed. Inappropriate interpolation may have been a problem in at least one study as Murray et al. (1979) have commented that incorrectly estimated norms are an alternative explanation for a South Carolina study in which actual gains were greater in fall-spring testing than in spring-spring testing (ESEA Title I Annual Evaluation Report: FY 1975, Office of Federal Programs, South Carolina Department of Education, Columbia, S.C., November, 1975).

DeVito and Long compared the effects of spring-spring vs fall-spring testing of educationally disadvantaged students on evaluation results. They found significant declines in percentile rank between spring and fall testing in most elementary grades (e.g., second graders: spring 33rd percentile, fall 12th percentile; fourth graders: 20th vs 11th percentiles). Unfortunately, DeVito and Long did not hold test form and level constant in all cases. Differences in content between levels could easily account for these results.

Faddis and Estes (1978) compared treatment effects for fall-fall and fall-spring evaluation cycles using Model B. No treatment effects were found for either cycle nor were there differences between results depending on the cycle. However, the authors cautioned that these results are only suggestive since a high attrition rate was apparent, interpolated norms were used for fall norms, and summer losses seen in other studies were not apparent.

In addition to the question of summer vacation declines is the issue of whether evaluation should concern itself with only the school year or with long-term effectiveness. David and Pelavin (1978) maintain that a program should demonstrate sustaining effects over the summer. They demonstrated that summer losses can be significant and recommended that evaluation be done on a fall-fall basis so that program effects over the summer can be ensured. However, mere evaluation on a yearly basis does not ensure long-lasting program effects and, as evidenced by Faddis and Estes (1978), high attrition may preclude evaluating on a yearly basis.

In summary, different evaluation periods may produce different estimates of treatment gain. This is due to two factors: use of interpolated fall norms and possible differential summer loss or gain

between treatment and control groups. It appears as though summer losses can be significant and linearly interpolated fall norms may not reflect this loss.

Two recommendations can be made. The User's Guide should specify norm interpolation method, since different methods yield different results. Secondly, as suggested by Murray et al. (1979), fall-fall, spring-spring, and fall-spring test data should be separated. Differential summer growth effects would then be easier to detect.

Test Administration, Scoring, and Analysis

Specified procedures are supposed to be followed in administering all standardized tests. These procedures may include a wide variety of features such as specific wording on various items, basal and ceiling level detection, timing, and the use of practice items. Testing conditions (e.g., quiet, well-lit room) are usually specified. Often, it is recommended that the test should be given only by those individuals trained in its administration. Deviation from the instructions can seriously alter students' test scores. This is particularly a problem with Model A, since scores must be comparable to the norm data (Horst & Wood, 1978). For Models B and C, testing conditions, including setting and time of testing, procedures, and administration must be comparable between treatment and control groups (Tallmadge & Roberts, 1978).

Following test administration, the dilemma of scoring the tests occurs. Usually, scoring is not done by hand but by a computerized scoring service. Following raw score determination, gains must be converted into NCEs. The conversion for each score requires several steps, usually including conversion to the publisher's standard score, then to the national percentile ranking, and finally to an NCE. If a

non-normed test was used, additional calculations are necessary. At each conversion step, errors in calculation and transcription may occur. While the scores and conversions performed by scoring services are more accurate than manually tallied and converted scores, the cost of such services can be substantial.

Analyzing the data, i.e., determining the effectiveness of an individual project, involves several steps. Each model prescribes a particular analysis technique, which can range from simple (Model A) to complex (Model B and C). Following analysis at the LEA, evaluations are passed on to the SEA which aggregates the data.

Several authors have discussed what may happen when appropriate procedures are not followed. Johnson and Thomas (1979) listed problems that can occur prior to scoring including lost answer sheets, improperly coded answer sheets, and matching pre- to posttests. Stonehill and English (1979) specified three types of errors which can occur: arithmetic (e.g., incorrect computations), procedural (e.g., use of inappropriate norms tables), and clerical (e.g., incorrect data transcription). They concluded these errors result in overestimates of program effect, since positive gains are rarely as well scrutinized as null or negative effects. Finally, they estimated that more than 95% of districts not using computer scoring and data processing will make some errors.

Stonehill and English call for greater reliance on computerized processing systems. However, Taylor (1981) commented that these also have drawbacks. Encoding of answer sheets is rarely checked, so that correct answers are marked incorrectly. Miscellaneous accidental marks on answer sheets, which do not affect hand scoring, may affect computer scoring.

Finally, when computerized scoring and analysis is performed, teachers are further removed from the evaluation and less involved in TIERS.

In conclusion, there appears to be no panacea which will eliminate errors in test administration and data collection. Unfortunately, errors in these areas can substantially alter the outcome of the evaluation. Training teachers in the whys and hows of evaluation procedures may contribute substantially to a reduction in error, but unfortunately imposes duties upon those who may have the least time available.

Summary

In summary, the general assumptions for all models have two basic types of problems, both issues of quality control. Technical problems are associated with test selection and evaluation model implementation, while processing errors occur in the scoring and calculation of results. It appears that the majority of these problems are due to two factors: (1) the flexibility of the system, and (2) the number of steps required to complete a given evaluation, i.e., the greater the number of steps, the greater the number of errors.

The problems of the system are caused in part by the mandate that TIERS be adaptable to any Title I project. Since the curriculum focus of a project is decided by the LEA/SEA, the evaluation system must be able to be used with any curriculum. Despite assistance given to the LEA/SEA by a Technical Assistance Center (TAC), these problems must in the end be weighed by the local evaluators. Inappropriate test selection can result in an invalid evaluation.

Unfortunately, it may be difficult at the LEA/SEA level to determine if the evaluation is valid. The fit of a curriculum to a test may not be good, resulting in no measured gain when gain existed; or may be overly specific, resulting in gains for too narrow an area which would not be reflected in improvement in the basic skill area. Floor and ceiling

that without treatment, the students would have the same percentile ranking on the posttest as they did on the pretest. Any percentile change therefore is attributed to the Title I project.

There are two general difficulties with the equipercentile assumption. First, growth rates of the publisher's norming population may be greater than those of the Title I participants. In other words, Title I students may "lose ground" over time resulting in a lower no-treatment effect than predicted by the equipercentile assumption. Secondly, the norming population may not be similar to the Title I population in terms of such factors as minority composition, age, or SES.

Growth Rates

There are several reasons why maturation rates may change over grade levels. Kaskowitz and Norwood (1978) examined gains necessary for normal growth on the MAT and compared them with gains necessary for educationally significant growth (defined as increase over normal growth by greater than 1/3 of a standard deviation). The rate of normal growth decreases over grade level, but the standard score standard deviations increase over grade level. Thus, 1/3 of an SD at grade 1 is a much smaller proportion of a year's growth than 1/3 of a SD at grade 7. Thus, educationally significant growth is more difficult to attain at higher grade levels.

Virtually no literature examines this problem coupled with functional (out of level) testing. For example, a fifth grader tested at third grade level would be required to perform at a third grade growth level which is greater than that for fifth graders since gains necessary for normal growth decrease over grades. However, educationally significant gains are less for third grade than for fifth grade, so while required growth may be greater, less is required for educational significance. In recommending

functional level testing for Model A, TIERS must assume that these increases and decreases balance themselves out. But there is no literature to support this unstated assumption. In other words, the assumption is that a child functioning at the third grade level is a third grader, in spite of age or previous difficulties in academic gain.

Cross-sectional Studies on the Equipercentile Assumption

The composition of a group of Title I students may differ from that of the test publisher's norming population in a variety of ways. The most obvious differences which may be related to academic achievement are SES, age, and race. How important these factors are with regard to TIERS is still a question, but RMC does caution the evaluator to use norms from a "comparable" population. However, "comparable" is never defined, and no guidelines are given for making this determination.

Two obvious problems in establishing comparability are (a) when Title I students are included in the norming population which would violate the assumption of no-treatment effect, or (b) if the Title I population is vastly different than the norming population. For example, Doherty (undated), as reported in Murray et al. (1979), stated that using norms not based specifically on disadvantaged, low achieving students gives non-valid measures of growth. To properly evaluate a group of students totally unlike those in national norms, restandardization on the appropriate population must be performed.

Mayeske and Beaton (1975) attempted to relate background and school variables to achievement. Using cross-sectional data, they determined the percentage of various minority students placing above selected percentile ranks for white students. For blacks, these percentages decreased over grade level, while for other minority groups, the percentages increased

over grade level. While these data suggest that groups may change their relative position over time in different directions than whites (thus invalidating the equipercentile assumption in some cases), the data are cross-sectional and may reflect population changes and differential drop-out rates rather than actual changes over time. Longitudinal data collection is usually necessary if conclusions regarding racial/SES differences are to be drawn.

Van Hove, Coleman, Rabban and Karweit (1970) examined percentile ranks of achievement test results for 7 cities at grades 6 and either grade 3 or 4. Linn (1978) converted the global results reported by Van Hove et al. into NCE scores. With the exception of one city, NCE scores dropped between lower and upper grades for nearly all minority schools. The same effect was seen in 5 out of 7 cities' nearly all majority schools. Excluding the gains seen in 3 cities (difference between lower and upper grades: .5 to 4.4 NCEs) the decreases for the nearly all minority schools (-.7 to -7.7 NCEs difference) were greater than the decreases for the nearly all majority schools (-.5 to -4.8 NCEs difference). These studies raise doubts about the usage of the constant NCE as a no treatment expectation, even though the data are cross-sectional, and the level of test was not constant across grades.

Coleman, Campbell, Hobson, McPortland, Mood, Weinfeld, and York (1966) compared the status of various minorities with the highest scoring group (white, urban, northeast population) using standardized scores on the STEP. Computing the number of SD units that each group was below the highest scoring group for grades 6, 9 and 12, the authors demonstrated that blacks and Puerto Ricans' scores tended to fall on verbal ability and reading relative to the highest group (from .1 to .5 SD units), but rise on

math tests (.1 to .5 SD units). Once again, the cross-sectional nature of the data makes it difficult to draw firm conclusions. However, all of these studies suggest a pattern of growth contrary to the equipercentile assumption.

Longitudinal Studies on the Equipercentile Assumption

Longitudinal studies are perhaps the best tests of the equipercentile assumption. Since the same group is monitored over time, variance due to true differences between populations, seen in cross-sectional studies is eliminated. Additionally, the need to ensure identical testing situations and identical race/SES composition is removed. Studies which have longitudinally examined the equipercentile assumption have concluded that it is valid in some cases, but not others (Powell, Schmidt, & Raffeld, 1979; Kaskowitz & Norwood, 1977; Hiscox & Owen, 1978; Armor, Conry-Osequera, Cox, King, McDonnell, Pascal, Pauly, & Zellman, 1976). All of these studies traced populations of students over at least one grade level to examine the assumption. Most have methodological problems which dictate caution in drawing conclusions from the data.

Part of the Armor et al. (1976) study traced achievement of over 700 predominantly minority program students over 4 years. The results show a percentile increase between grades 3 and 4 but steady decreases thereafter (33rd percentile in grade 4 to 29th percentile in grade 6). The increase in the first year was attributed by the authors to a change in tests. However, the results of this study are confounded by several factors. First, the equipercentile assumption is assumed only in the absence of special programs, but most of the students were in special programs. Since this study, overall, was an evaluation of the Preferred Reading Program, it is possible that treatment damaged the students or that the

tests used did not properly match the curriculum. Secondly, there is no indication of whether the same level of test was used for each pre- and posttest cycle. If this is not the case, the differences can be attributed to a change in test levels.

Problems with different test levels and time of testing are sources of difficulties in a study done by Kaskowitz and Norwood (1977). The study hypothesized that the equipercentile assumption may not adequately describe Title I type student performance in the absence of a special program. While it was demonstrated that such a population showed percentile decreases over 3 years with respect to the norm group, testing occurred at differing levels and off publishers' norming dates. The authors concluded that norms based on the standardization group will be too high for an educationally disadvantaged population.

Powell et al. (1979) analyzed pre-posttest scores on the MAT in a one-year spring-spring testing evaluation. Three groups of pre-posttest cycles were examined longitudinally: (1) end of second grade - end of third grade, (2) end of fourth grade - end of fifth grade, and (3) end of second grade - end of fourth grade. Although the assumption of equipercentile growth held for the reading subtest results, the standard scores on the math subtest were statistically significantly different at the .05 level from those expected using the equipercentile assumption. Whether this can be transformed into similar differences in terms of NCEs is undetermined. The main difficulty in interpretating these data is, again, problems with the test level utilized. In all of the cycles examined, the form and level of the test changes from pre- to posttesting. As mentioned previously, differences found when using different test levels may result from differences in content. While the evaluators

appear to have assumed equivalency between test forms/level, probably because the MAT was used throughout, equivalency was not established. Differences in content and better curricula-test matching could easily account for the results.

Hiscox and Owen (1978) also attempted to longitudinally analyze data to answer questions regarding the equipercentile assumption. In this study, the authors carefully examined students' achievement test scores and percentile ranking over 4 years for both Title I and comparable non-Title I students. As pointed out by the authors, the problems encountered in the study make interpretation difficult. First, attrition over the 4-year period was substantial. At the high school level, for example, more than two-thirds of the students were "lost" by the fourth year. Secondly, out-of-level testing was widespread and the effect of this on expanded standard scores or NCEs is not known. While several groups show enough change over a 3-year period to question the tenability of equipercentile assumption, the authors were unable to determine whether real differences in achievement, changes in test levels over the 4 years or the lack of complete data caused the percentile change. In other words, whether the model (i.e., the equipercentile assumption) or the problems with the data are at fault was undetermined.

Summary

In summarizing the literature on the equipercentile growth assumption, very few positive statements can be made since methodological problems abound. Instead of collecting their own data, most studies simply used data from previous testings. Consequently, test level for pre- and posttesting is rarely constant. Without the same test being used for pre- and posttesting, it is difficult to determine whether any changes

are due to the level of test changing, or to true achievement levels. Additionally, there is no control for equivalency among testing conditions. Secondly, it is difficult to determine if standardized norms are applicable for the group being examined: out-of-level testing would invalidate the norms due to age and maturation effects, racial and SES composition may be drastically different, and testing may not have occurred on/near publisher's norms (Noggle, 1977).

Finally, there is seldom any standard for defining a "Title I eligible student". Comparing studies is difficult in this light, especially since several authors have suggested that educationally disadvantaged students may progress at different rates (Faddis & Estes, 1978; Hiscox & Owen, 1978; Mayeske & Beaton, 1975; Van Hove et al., 1970). By having no standard of performance for Title I eligibility, severely educationally disadvantaged students may be aggregated with mildly educationally disadvantaged students. Test norms would not be appropriate for both populations since rates of growth for the severely educationally disadvantaged would be slower than for mildly educationally disadvantaged students.

The appropriate study on the equipercentile assumption for Title I students has yet to be performed. Such a study would require pre- and posttesting students selected on the basis of a selection test as educationally disadvantaged. Test levels should be identical for pre- and posttests, and no Title I type intervention should occur. If this population demonstrated no declines or gains in percentile ranking, then the equipercentile assumption could be said to hold for Title I type students.

Until such a study is performed, use of Model A must be coupled with strict controls. Whether or not equipercentile growth occurred in the past may be indicative of current growth. As suggested by Murray et al. (1979), the evaluator could examine this in one of three ways. Pre Title I records could be checked for students who would have been selected for Title I, and, if over a similar period of time this sample maintained their percentile status and the composition of the sample is similar to the current Title I sample, the equipercentile assumption may obtain. If such data is not available, the percentile status of the district as a whole could be traced over time. If it was maintained, it would also lend support for the equipercentile assumption within this population, although this support would not be as strong as with the previous method. Finally, if no historical data was available, a current local group of students similar to those in Title I could be examined for equipercentile growth. This is the weakest method for eliciting support for the equipercentile assumption.

In summary, without local evidence in support of the equipercentile growth assumption, Model A should not be used. Without such evidence, it would be difficult to make valid conclusions regarding the data.

Selection Testing

In order to be a participant in a program to be evaluated using Model A, selection must be based on a test which is not used as the pretest. If selection is based on a test given prior to the pretest, regression towards the mean is expected to occur between the selection and the pretest, and not between pre- and posttesting. However, regression is directly tied to the amount of correlation between two tests--the lower the correlation, the greater the regression. Over time, correlation

between pre- and posttest drops, thus increasing the regression that will occur between these pre- and posttests, even if regression has already occurred between selection and pretest (Glass, 1978; Burton, 1978). Such regression will result in an artificially high estimate of gains due to academic achievement. Campbell and Stanley (1966) caution against use of this type of procedure due to such "pseudo gains". While Yap (1978) demonstrated that selection tests could be used as pretests, the study was limited since it assumed a normal test score distribution. Real data may not be normally distributed. In summary, the validity of the assumption that regression occurs between selection and pretest and no regression occurs between pre- and posttest is unsupported by the literature.

Date of Testing

For the standardized norms to be appropriate when used to evaluate Title I students under Model A, the User's Guide stipulates that the actual date of testing occur within two weeks of the publisher's date of testing or six weeks if interpolated norms are used. Pre- and posttesting should be equally distant from these published dates, i.e., if the pretest is given four days prior to the published date, the posttest should also be given four days prior to the published date.

As discussed by Bridgeman (1978) and Baker and Williams (1978), several problems are encountered with interpolation. The foremost is the manner of interpolation. One method entails plotting of equal standard score lines on a graph of NCEs versus dates. Another would be to plot equal NCEs on a graph of standard score versus dates. These two methods will yield the same results only if the standard scores are normally

distributed within the group at both norming dates which would result in a linear relationship between standard score and NCEs (Baker & Williams, 1978). These authors recommend the former method on the basis that if the scores are not exactly normal, and given that rounding occurred during norm development, this method will yield more accurate results: it requires only one calculation whereas the latter method requires several calculations permitting more error to occur.

The second problem with interpolating norms is whether to interpolate or extrapolate. If testing should occur October and April, and instead occurred in September, one could interpolate using the previous year's April norms and the current year's October norms. However, as previously discussed, this type of norm will not account for summer growth or loss. Extrapolation from the current year's October and April norms is recommended (Baker & Williams, 1978).

Additional difficulties can occur when students are absent on the appropriate testing day. While make-up testing is mentioned infrequently in the literature, the problem of make-up testing and the equating of testing conditions may create additional problems in interpolation and interpretation. Finally, most of the literature pertaining to interpolation discusses the topic from the standpoint that it should only be done if there is no alternative. Indeed, careful advance planning in advance can eliminate the need to interpolate norms. However, as discussed previously, interpolation methods should be specified by TIERS, so that results are comparable across projects.

Summary

There are serious threats to the internal validity of Model A in terms of possible historical, maturational, selectional, and instrumentational differences between the norming sample and the Title I sample. An additional threat comes from the assumption that statistical regression will occur only between selection and pretest. Finally, the use of interpolated norms can result in bias estimates of expected gains. The combination of these factors make Model A the weakest of the three models. Unfortunately, it is the most frequently used model as it is the simplest and least expensive to implement.

C. ASSUMPTIONS OF MODEL B

Two issues are important for determining the validity of Model B. The first deals with the appropriate control group, and the second deals with the appropriate statistical adjustment for non-equivalent control groups (Tallmadge & Horst, 1976). Model B is the strongest of the three models in that its design, i.e., the use of a control group, is well supported in the literature (e.g., Campbell & Stanley, 1966). The problems which occur with this model are not difficulties with the validity of the assumptions used, but difficulties with their implementation. Unfortunately, the use of a proper control group is extremely difficult to achieve due to ethical constraints as well as definition.

Composition of Control Group

Model B calls for comparing performance of Title I participants to a no-treatment control group. This control group must be selected using the same criteria used to select participants, and to receive the same educational advantages and curricula as Title I participants except the actual Title I program being evaluated. Implicit in this model are the factors of identical testing conditions, test levels, and racial/SES group composition. In short, all factors that could affect the performance of the Title I participants must be incorporated into the control group.

Unfortunately, this ideal control group is difficult to locate. For example, ethical considerations preclude random assignment into treatment and no-treatment groups. SEAs can circumvent this by having Title I and non-Title I schools, so that the control group can be drawn from the non-Title I school. This can create some problems since Title I schools are usually chosen on the basis of greatest need. In the relatively rare instances where students participate only half of the school year, the students in the program during the fall can be compared to those who did not participate in the fall but in the spring. In such cases, random assignment can occur in terms of who receives the program when.

Unless the control group is comparable to the treatment group, Model B cannot be used. If implementation occurs using similar but non-equivalent control groups, statistical adjustments are necessary. Unfortunately, the proper use of these adjustments is difficult to ascertain, either in the materials produced by RMC, or in the evaluation literature (Goldman & Crane, 1980). The following section will examine the literature pertaining to such adjustments.

Statistical Adjustments

Basically, two types of adjustment are available to the user of Model B. These adjustments are analysis of covariance and principal axis adjustment. Covariance analysis, which uses the slope of the common within-group posttest on pretest line to adjust for the initial differences, is most appropriate when used with groups which are "random in effect" (Tallmadge & Horst, 1976). Principal axis adjustment uses the ratio of the pooled within-group posttest standard deviation to the pooled within-group pretest standard deviation. Adjustment is then made by subtracting the control group's posttest mean from the product of the principal axis and the pretest difference between the groups.

Essentially, principal axis adjustment is analysis of covariance when the correlation between covariate and the dependent variable is set equal to 1.0. In other words, the use of principal axis is appropriate if the groups exhibit stable differences over time. Kenny (1975) argues that principal axis adjustment is usually most appropriate since those assigned to Title I are usually the most needy students so the control group represents a population drawn from a radically different environment. Differences, usually seen in terms of higher pretest means in the control group, may cause serious threats to the internal validity of the non-equivalent group design since the difference may be due to dissimilar maturation rates and these may interact with selection (Campbell & Stanley, 1966).

The use of the principal axis adjustment is dependent upon stable differences between the groups being measured. This assumption is based on the fan-spread hypothesis which states that the difference between group means is constant over time relative to the pooled standard

deviation within groups (Kenny, 1975). A major difference between this and the equipercentile assumption is that with the fan-spread hypothesis, performance of both groups is being measured while with the equipercentile assumption, the performance of a population which may not be included in the norming sample is being measured against that norming sample. If the population being tested was in fact a subsample of the norming population, the equipercentile assumption and the fan-spread hypothesis would be synonymous.

Several authors have maintained that the fan-spread hypothesis has serious weaknesses (Linn, 1978; Linn & Werts, 1977; Goldman & Crane, 1980). Linn and Werts (1977) demonstrated that if the fan-spread hypothesis is untrue, standardized-gain-score methods can lead to biased estimates of gain. Goldman and Crane (1980) used computer simulated data to examine the bias which results when different analytical techniques are used in different situations. With regard to the principal axis adjustment, the four conditions important to this review were:

1. Random assignment, equal principal axes.
2. Nonrandom assignment, equal principal axes, pretest means and SDs unequal.
3. Nonrandom assignment, unequal principal axes.
4. Nonrandom assignment, unequal principal axes, equal pretest means.

Three scores were obtained for each condition: unadjusted, covariance, and principal axis. Unequal principal axes violate the fan-spread hypothesis since a common within-group principal axis should not be calculated.

As predicted, any adjustment--either covariance or principal axis--is better than no adjustment at all. For the ideal situation (condition 1) all three methods demonstrate a small but equal amount of bias (-.26 NCEs unadjusted, -.16 NCEs covariance, and -.13 NCEs principal axis). For condition 2, the ideal situation for use of principal axis, principal axis produced the least bias estimate, differing by only .07 standard score units from the actual gain. In condition 3 where unequal principal axes violate the fan-spread hypothesis, principal axes produced less bias than covariance (-4.19 NCEs vs -5.80 NCEs, principal axis vs covariance respectively), but substantial bias was present. In condition 4, where again the fan-spread hypothesis is violated, covariance and principal axis produced about the same amount of bias (-4.44 NCEs and -3.97 NCEs respectively.

In summary, bias was least where principal axes of the groups were equal. The primary implication is that the principal axis adjustment is best in situations where the fan-spread hypothesis is operating. In any case, the principal axes method produces the least bias, although as conditions degenerate, greater bias appears. Finally, all biases were in negative directions, which would lower treatment effect estimates. The authors recommended that despite difficulties in discerning whether the fan-spread hypothesis is in effect, the principal axis method is best in all non-equivalent group situations, since it produces the least bias and any treatment effect seen will reflect the minimum gains made, since the bias is in the negative direction.

Summary

It appears as though the problems of Model B are somewhat more surmountable than the problems associated with Model A. The control group design has a strong background of support in the evaluation community. However, it is crucial that the data be examined for equivalency between control and treatment groups. If non-equivalencies exist, differences must be adjusted for using the appropriate method. Although not well discussed in the literature, large attrition can affect either statistical adjustment. Although use of the principal axis adjustment assumes that the fan-spread hypothesis is in effect, violation of the fan-spread hypothesis does not appear to have such drastic effects that evaluations would be invalid. However, in studies where possible gains are unknown or where they may be small, the evaluator is cautioned against the use of principal axis if there is any question of violations of the fan-spread hypothesis (e.g., unequal principal axes), since its use may lower treatment effect estimates thereby overshadowing real gains.

D. ASSUMPTIONS OF MODEL C

Model C is specifically designed not to have an equivalent control group. A comparison group is formed by including all students scoring above a specified score on the pretest, while all those who scored below this score are included in the treatment group. Post on pretest regression lines are fitted to both groups. If there was no treatment effect, the Title I regression line would be a downward extension of the comparison group's regression line. If the treatment was effective, then the regression line for the Title I group would lie above and parallel to

the regression line for the non-Title I group. Two assumptions must be met for this model to function properly. The first is firm adherence to the cutoff score. The second is that both regression lines are parallel and linear.

Adherence to Cutoff Scores

The use of a distinct cutoff score is required in Model C to distinguish between the control and treatment groups. Unfortunately, this is probably the most common problem encountered during Model C implementation. Despite the requirement of a strict cutoff score to separate Title I program participants from the comparison group, many programs do not adhere to the score (Yap, Estes, & Hansen, 1979). There is often a gray area such that the single cutoff score becomes a band of cutoff scores.

Yap et al. examined the various evaluation outcomes which could occur when a band of cutoff scores is used. These conditions included: (1) adherence to the strict cutoff as outlined by Tallmadge and Wood (1976); (2) use of a band of cutoff scores, but students in the band excluded from analysis; (3) use of a band of cutoff scores with those in the band randomly assigned to treatment or comparison group and included in the analysis, and (4) use of a band of cutoff scores with students in the band assigned to treatment or comparison group on the basis of teacher ratings and included in analysis.

Using computer simulated data, they found that when a strict cutoff was used, relatively unbiased estimates of effects resulted. In the second condition, where students in the gray area were excluded from analysis, only slight differences between estimated and actual gains

appeared (all less than 2.0 NCEs). When students in the gray area were randomly assigned to groups, practically no bias was introduced. However, in the final case where teacher ratings were incorporated into placement, several sources of bias resulted. For almost half of the instances, the difference between estimated and actual gains was greater than 1.0 NCE and reached 3.36 NCE in one case. Bias tended to increase as the width of the gray area increased and as the number of students in this area increased. In general, bias for all conditions, when it occurred, tended to favor the treatment group in that the estimated gains were higher than the actual gains. However, the use of teacher ratings tended to suppress actual treatment effects, yielding an estimated gain which was less than actual gain.

One difficulty with the Yap et al. study is that the data were simulated on the basis of equal growth rates across all students. As was seen in the equipercentile growth assumption of Model A, lower functioning students do not necessarily have the same growth rates as higher functioning students. However, these results do suggest guidelines to be followed if additional variables are used in selecting students for Title I programs. If students who do not meet a strict cutoff on the selection measure are permitted to enroll in Title I programs, their scores should not be included in the analysis when the program is evaluated.

Parallel and Linear Regression Lines

Model C is based upon the assumption that regression lines for treatment and control groups will be parallel and linear. The User's Guide states that a test which produces curvilinear regression should not

be used. However, it may be difficult to determine beforehand whether curvilinearity exists with a particular group. As mentioned briefly in the section on out-of-level testing, floor and ceiling effects tend to result in curvilinear regression lines (Estes & Anderson, 1978). Additionally, parallel regression lines are based on the assumption of equal maturation rates between the two groups. Unlike Model A, this assumption is not usually violated since all students--comparison and treatment--are drawn from the same local population, whereas in Model A, a subset of the local population is compared to a national standard.

Unfortunately, even if the evaluator adheres strictly to the User's Guide, there are many instances where non-linearities could appear. The few articles which discuss nonlinearity are attempts to make adjustments so that nonlinear data can still be used.

Estes and Anderson (1978) analyzed the pre-post test scores of 730 ninth graders on three math tests (Comprehensive Tests of Basic Skills Math Subtest (CTBS), Shaw-Hiehle Individualized Computational Skills Test, and the Minimal Mathematics Proficiency Test (MMPT) to test the no-treatment expectation for Model C. Hypothetical treatment and control groups were formed and analyzed as dictated by Model C. In spite of no treatment, estimates of treatment impact in NCEs at pretest-mean and cutoff score were 4.4 and 1.3 for the CTBS, -6.30 and -4.42 for the MMPT and 2.15 and 2.90 for the Shaw-Hiehle. Statistical tests comparing "treatment" and "control" groups' results were statistically significant at the .05 level in almost all cases.

Floor effects were detected on the CTBS pretest and ceiling effects on the MMPT posttest. These effects were demonstrated by unequal treatment effect estimates at the pretest mean and cutoff score. The

authors recommend that if floor or ceiling effects are present, then these students should be deleted from the evaluation. Unfortunately, deleting such students may alter the composition of the population in terms of growth since specifically deleting higher level students or lower level students is not random deletion..

Echternacht and Swinton (1979) propose four possible solutions to the curvilinearity problem. These include Mosteller and Tukey's re-expression of posttest scores, differential weighting of scores in different parts of the pretest, use of quadratic regression lines and extrapolation to obtain the no-treatment expectation, and the use of fitting parallel lines to the data for each group. Each method has drawbacks.

Mosteller and Tukey's re-expression of the posttest scores is a rough method and is described as "as much of an art as a science". If few data points exist, results can vary depending on which points are selected for re-expression, which may result in unreasonable expectations.

If a computer is available to the evaluator, the technique of weighting scores may be applied. However, Echternacht and Swinton concluded that such weighting produces essentially the same regression function as one would achieve using quadratic fits. If the data are nearly linear such fits may work well, but when there are few data points, or if an unusual function exists, fitting or extrapolating from the fit can be dangerous.

Finally, Echternacht and Swinton discussed the use of parallel lines fit, i.e., analysis of covariance. The traditional Model C approach differs from the parallel lines fit in that the traditional approach regresses the data from the control group and extrapolates to the treatment group whereas parallel slope fits all of the data on the

assumption of parallel slopes. When ceiling effects are present, parallel slope fit is better than the Model C approach. When floor effects are present, the Model C approach is better. If a treatment x pretest interaction occurs, Model C may be better, since the parallel lines fit will confound such an interaction with the estimate of treatment impact. However, if no such interaction is present, parallel lines fit is better provided the control group is not greatly larger than the treatment group (Cochran, 1969). Echternacht and Swinton concluded that while Model C works well in the absence of floor and ceiling effects, when such non-linearities are present the data should be fitted a variety of ways and results compared. If fits with the parallel lines procedure and Model C procedures provide similar results, the curvilinearity is probably not serious.

Summary

In terms of rigor, Model C falls between Models A and B. As in Model B, the problems encountered appear to be surmountable. However, for Model C to produce valid results, evaluators should firmly adhere to the cutoff score. Special attention should be given to floor-ceiling test effects which create non-linear regression lines. When non-linearity is unavoidable, the data should be fitted according to procedures outlined in Echternacht and Swinton (1979).

THE COMPARABILITY OF MODELS

According to the User's Guide, the use of any of the models will yield comparable results. This aspect is extremely important since results are intended for aggregation at SEA and national levels. If the models are not comparable, this aggregation will produce inappropriate/faulty summaries regarding the impact of Title I. Several studies have addressed the issue of comparability. While a portion of this research has been discussed in the previous section with regard to the use of normed versus non-normed tests within a given model, the following section will detail various field studies and computer simulations that have directly compared the results obtained from different models.

To properly compare different models, the assumptions applicable to all models plus the specific assumptions for each model must be followed. For example, to compare Model A to Model B or C, all students being compared must: (1) be selected on the basis of a selection which does not serve as the pretest, and (2) must be tested within six weeks of publisher's norm dates in addition to meeting the requirements for Model B (control groups only randomly different and appropriate analysis used to adjust pretest scores) and/or Model C (control group is all students scoring above a specified criterion on the pretest). In addition, assumptions concerning appropriate test selection and administration must be met. In almost every study purporting to compare two models, major assumptions are violated.

One of the major difficulties in making these types of comparisons is that the data cannot usually be examined post hoc. Since the decision to use a particular model is usually made beforehand, and a different model's requirements are frequently not met. Table 1 lists those studies which report comparisons and which, if any, assumptions were violated.

Table 1

Summary of TIER'S Model Comparison Studies

STUDY AND YEAR	MODELS COMPARED	ESTIMATE OF TREATMENT IMPACT BETWEEN MODELS	COMMENTS	CRITICAL REVIEW
Crane & Coch, 1979	A_1 vs B_1	A = B (Kindergarten) B > A (1st & 2nd grade)	1) Selection for participation Not Random 2) Selection on basis of either "Review of previous testing" or "recommendation by Child Study Team" 3) No Selection Test (Model A required), Interpolated norms 4) Kindergarten 1st & 2nd grades examined, RMC only intended for 2nd-grade and above	Procedure not specific enough to tell exactly what happended, however, there are sufficient violations of the models to place any conclusions in doubt
Haddis, Arter, & Zwertchek, 1979	B vs A (w/local norms) vs A (publisher's norms)	For READING Model B showed less impact than Model A with national or local norms (-.5 vs -1.8 vs -5.4) For LANGUAGE A national < B < A local (-1.4 < -3.5 < +4.8)	1) Interpolated norms, Testing off publishers dates (Model A required), Control groups not randomly different (Model B required) 2) Attrition problems different and greater in Title I schools versus control schools. 3) Publishers test norm included Title I students, therefore, greater growth at low level achievement included. 4) Control group formed to check equipercentile growth for Model A showed percentile losses similar to Title I students, i.e., percentile not maintained	Too many violations of Model A to permit true comparison (i.e., interpolated norms due to testing off date, equipercentile not maintained in control group
Gabriel, Stenner, & Troy, 1977	A vs. B	NCEs 1st Grade A B A < B Reading 3.9 vs 5.5 A = B Math 6.2 vs 6.7 2nd Grade: A > B Reading 9.5 vs -2.8 ns Math 8.1 vs .09 3rd Grade A = B Reading 3.1 vs 2.3 A > B Math 7.9 vs 3.0	1) Interpolated norms in grades 2 & 3 may have caused differences between Models A and B. Interpolated norms not used in 1st grade 2) Authors speculate that removal of low achievers helps rest of class 3) Different achievement tests were used for each model (NRT for A, CRT for B) 4) Curriculum not specified	Use of interpolated norms may have caused differences seen in 2nd & 3rd grades. These were not used in 1st grade and no differences seen
Hardy, 1979 (as reported by Murray et al 1979 and Echternacht, 1980)	A vs C	A > C (size of differences unknown)	1) Most Model A tested fall, spring while Model C tested spring, spring 2) Aggregated across schools and grades but similar school sizes and close pretest averages.	Summer losses could have easily caused lower Model C estimate of treatment impact
House, 1979	A vs C	Reading, Grade 4 C Slightly higher estimate than A (A = -.2 NCE, C = .8 .5 NCE). Grade 6, 8 A Slightly higher estimate than C (A = 2.3 (grade 6), 4.1 (grade 8), C = 1.4 (grade 6); 2.5 (grade 8)) Math Grade 4--C > A A= .6 C=.3 Grade 6--C = A 3.5 3.4 Grade 8--A > C 6.6 5.6	1) Difficult to determine if both models used same data base. 2) Sizes of treatment and control groups greatly different. In reading Grade 4: 1408 Treatment vs. 934 Control; Grade 6 1806 vs. 559; Grade 8: 1159 vs 877, whereas in math, reversed Grade 4, 685 vs 1106, Grade 6, 789 vs. 1032; Grade 8, 485 vs. 1044. 3) Slopes are different for Model C regression lines	Since Model C regression lines appear to be unadjusted for pretest scores and slopes were not parallel, it appears as though it was not implemented correctly. It is difficult to determine exactly what the procedure was.
Tallmadge & Wood, 1980	A vs B vs. C 3 different analysis (1) large heterogeneous sample, (2) smaller homogeneous sample (low SES; low achievers), (3) small sample (simulated project)	Model A tended to produce positive bias gain estimates of approximately 1 NCE. Model C tended to be less accurate than A & B, and most sensitive to disturbances with its assumptions (e.g., floor-ceiling effects). But, IN GENERAL, Models equivalent	1) Extreme care was taken not to violate the basic assumptions of any model	On the whole, this study has appropriate data to make its conclusions regarding overall model equivalency when careful control is implemented.

In examining Table 1 only one of the six studies meets the requirements for each model being compared: Tallmadge and Wood (1979). All of the remaining five studies (Crane & Cech, 1979; Faddis, Arter, & Zwertchek, 1979; Gabriel, Stennar, & Troy, 1977; Hardy, 1979; House, 1979) contain methodological flaws which could account for differences obtained. In two of these studies (Crane & Cech, 1979; House, 1979), methodology and procedure sections are not specific enough to make determinations about study validity.

Tallmadge and Wood (1979) conclude that the models are comparable, while the five remaining studies, albeit flawed, do not come to this conclusion. Obviously, more properly controlled studies need to be performed in this area. The current collection of literature, however, is informative from the standpoint of how models are actually implemented at the local level versus how RMC envisions their implementation (Tallmadge & Wood, 1979). If the models are not strictly implemented, it appears as though the "garbage in-garbage out" effect holds. This problem, in and of itself, may cast suspicion on the results of TIERS, not because of any problem within the system, but due to external factors.

SUMMARY AND CONCLUSIONS

The development of TIERS represents the most comprehensive attempt on the part of the federal government to regularly and objectively evaluate a federally-funded program of this magnitude. In examining the validity of TIERS, the fact that it is a relatively new endeavor must be kept in mind. This review has examined the literature which assesses TIERS to determine if the system is and/or could be valid as an evaluation system.

Three types of concerns are discussed in the literature. These include: whether the statistical assumptions of the models are valid, whether proper implementation of a given model can occur at the local level, and whether the models are comparable. This section will examine these concerns and suggest possible solutions beyond the recommendations made in the review.

Validity of Statistical Assumptions

The first assumption of Model A is the equipercentile growth assumption. It is assumed that without treatment, students will maintain their pretest score percentile ranking on the posttest. There are serious threats to the internal validity of Model A. These are due to possible differences between the norming sample and the Title I sample in the areas of maturation, selection, and instrumentation. The second assumption of Model A is the regression effect. It is assumed that statistical regression will occur only between selection and pretest and not between pre- and posttests. The literature does not support this assumption.

The fan-spread hypothesis is the primary assumption of Model B. This assumes that the difference between group means is constant over time relative to the pooled standard deviation within groups. This assumption is only in effect when the principal axis method of adjusting pretest scores is utilized, i.e., when the differences between treatment and control groups are presumed to be fixed. Demonstrations using simulated data have shown that violation of the fan-spread hypothesis will not drastically affect the estimate of treatment gains.

Model C assumes parallel and linear regression lines. Parallel lines follow from an assumption of equal growth rates in comparison and treatment groups. Curvilinear lines will result if floor or ceiling test effects occur. Curvilinear regression lines can be fitted according to various procedures outlined in Echternacht and Swinton (1979).

Models B and C have the strongest support in the literature, while Model A appears to have substantial difficulties. With regard to all models, the major difficulty apparent in TIERS is that no provisions are made for what to do if an assumption has been violated. It is probable that the typical evaluator will proceed regardless of whether an assumption has been knowingly violated. There are no provisions for salvaging violated data and still obtaining meaningful information.

Proper Implementation

Proper implementation of any model requires a number of decisions to be made. These include: what test to use, when to administer it, how to analyze the data, which model to use, and how to select participants and controls. Errors made in proper implementation fall into two categories: technical errors and processing errors.

Technical errors tend to occur when user guidelines are not properly implemented. They can occur in several areas including test selection, date of testing, data analysis, choice of a model, and selection of participants and controls. These errors may be very hard to detect since the whys of specific decisions in these areas may not be known.

Processing errors occur during scoring and calculation of the results. These errors are difficult to detect if they result in positive treatment gains, since such results are expected and do not lead to double checking of the data. Negative results due to processing errors tend to be corrected.

TIERS has no way of forcing proper implementation. Despite the step-by-step nature of the User's Guide, many areas, such as matching the test to the curriculum, are not fully covered. Technical papers prepared in conjunction with the User's Guide, while they can be helpful, may also be overlooked.

Model Comparability

Unfortunately, most of the literature in this area violates various model assumptions, thus rendering the comparison useless. The excellent study by Tallmadge and Wood (1979) does conclude that the models are comparable. However, the remaining studies, albeit flawed, do not.

Unfortunately, it is these latter studies that are representative of TIERS implementation at the local level. These studies violate a wide range of guidelines established by TIERS or by other evaluation literature. If these types of violation are occurring at the local level, then it is difficult to evaluate the results of a project or to gather comparable data.

Possible Solutions

Due to the lack of support for its assumptions regarding equipercentile growth and regression, Model A appears to be inappropriate for Title I evaluation. Models B and C are more difficult to implement than Model A and are more costly, since at least twice as much testing is required due to the additional control/comparison groups. However,

Models B and C, if properly implemented, appear to evaluate Title I projects more effectively. However, if the existing literature is any indication of what happens at the local level, proper implementation of any of the models is not occurring. Hence, conclusions based on TIERS must be viewed with caution. Particularly for Models B and C, the suggestions made throughout this review regarding proper implementation, if used, would certainly result in better evaluations.

One solution in improving TIERS and its implementation is the use of well-trained evaluators at all levels of Title I program evaluation. Currently, TACs oversee the evaluation system, give workshops in its usage, and consult with LEAs and SEAs on implementation of TIERS. However, much of the existing literature demonstrates that these efforts have not been extensive enough, as the literature reporting on TIERS contain a wide variety of violations of the system. It seems as though those who are responsible for the implementation of TIERS at the school level are faced with many of the decisions required by the system but lack the expertise to make these decisions.

Murray et al. (1979) have suggested the area of evaluation is not at a point where valid results and conclusions can be made by those who are inexperienced. It is probably a small minority of those implementing TIERS at the local level who understand the consequences of violating the assumptions of the models or could recognize such violations. Title I and Title I evaluation are very costly, and it seems inappropriate that conclusions regarding Title I are drawn by inexperienced personnel.

Unfortunately, further training of school personnel or adding personnel specifically trained in evaluation would be costly. Perhaps, it is time to examine what is really needed in terms of evaluating Title I.

The purpose of Title I evaluation is twofold. First, there is accountability for monies spent on Title I so that such expenditure can be justified. Secondly, there is information provided to the LEA and SEA regarding whether their projects increase the basic skills of the students selected for participation. Currently, it is suggested that TIERS be used to evaluate every Title I program every year.

Title I involves thousands of programs and millions of students. Many of these programs do not change year to year. The same curriculum is used over again provided some indication exists that it is effective. Changing the curriculum year to year would be very expensive. The question arises as to whether it is necessary to evaluate the same curriculum every year, considering the composition of the local population is often constant, and Title I participants are similar year to year. In terms of the purpose of Title I evaluation, solutions not requiring yearly evaluation should be examined.

Three possible solutions exist which do not call for yearly evaluation. The most obvious is to stagger evaluation so that all programs are evaluated every other year, or every third or fifth year. A second solution is to evaluate only new curricula. Once a new curriculum is evaluated in a given school and shown to be effective, it could continue indefinitely without reevaluation at that site. A third is to randomly sample Title I programs for periodic evaluation (every 1, 2, 3, or 5 years). This evaluation could be conducted by highly trained evaluators with the technical expertise necessary to implement TIERS correctly. This study would provide the information necessary to justify the program at the natural level. LEA's would then be free to

use whatever evaluation procedures they want to make local programmatic and curriculum decisions. Of course some combination of these three options is also necessary.

The basic idea behind all of the solutions is that when evaluation occurs, it should be correctly implemented. Evaluation year after year of the same program is neither cost effective nor will it yield useful information unless proper implementation occurs. Non-annual evaluation would cost less than training current local evaluators in proper model implementation or hiring additional personnel trained in evaluation. Non-annual evaluation would be done at the school level by non-school personnel experienced in evaluation. Teachers and/or administrators who are currently carrying a full work load would not be given the additional burden of Title I evaluation.

In conclusion, Models B and C of TIERS appear to yield useful information regarding the effectiveness of Title I. However, if the models are not properly implemented, (which appears to happen frequently) the results are not useful and may lead to incorrect conclusions regarding the effectiveness of a given project. A solution to this is to use personnel trained in evaluation at the school level to ensure proper implementation. Costs of such a solution may be prohibitive. For this reason, non-annual evaluations are recommended in addition to the use of evaluators at the local level. Implementation of this recommendation would be more cost effective and would result in better evaluations.

CHAPTER III

A COMPARISON OF MODELS A AND B

Procedures

For purposes of comparing the estimated impact of the same Title I program using Models A and B, the Title I program in the Salt Lake City School District in Salt Lake City, Utah was considered. The district's Title I program was being operated in eight schools. Seven additional schools in the district had been included in Title I previously or were being considered for potential expansion of the Title I program.

Data regarding each of the eight Title I schools and the seven potential comparison schools on poverty level, mobility, daily attendance, average IQ of student body, percent minority, and reading and math scores from the previous spring, were collected. This information is presented in Table 2. Based on an analysis of these data, two schools (Wasatch and Nibley) were dropped as potential comparison schools, leaving five schools to be used in the actual comparison of Models A and B. There were no statistically significant differences between any of the Title I or comparison schools on the spring achievement test data reported in Table 1.

Guidelines for implementing both models suggested by Tallmadge and Wood (1976) were followed. The guidelines for implementing Model A recommend that tests be administered within two weeks of the empirically established norm date, but allow up to six weeks if scores are extrapolated. In this case, the pretest was administered five weeks before the empirically established norming date. The posttest was administered two weeks before the empirically established norming date. In analyzing the data, adjustments were made by

Table 2
School Means on Various Factors Used in Selecting Comparison Schools

	School	Poverty 78	Poverty 79	Mobility	Attendance	IQ	Percent Minority	Achievement Test Raw Scores: Reading 2	Reading 3	Reading 4	Math 2	Math 3	Math 4
1.	Franklin	34.9	38.7	42.6	97.69	98.4	43	113.18	116.62	78.64	38.70	58.86	55.20
2.	Jackson	58.4	47.5	58.7	97.70	94.4	52	113.80	118.55	73.36	43.65	67.15	44.85
3.	Lincoln	65.8	66.5	60.6	95.75	95.4	43	101.49	114.28	76.67	40.16	64.85	47.74
4.	Lowell	68.5	38.1	62.0	98.18	104.7	15	128.33	134.10	91.10	47.42	71.36	59.58
5.	L. Benion	45.6	61.2	65.2	95.98	94.7	31	118.84	118.63	86.97	43.80	66.14	55.02
6.	Parkview	29.3	26.2	43.6	98.54	107.6	38	96.26	114.34	78.59	41.38	64.42	49.39
7.	Washington	52.7	49.0	72.3	98.35	94.9	34	115.54	126.59	82.57	47.15	71.72	54.13
8.	Whitman	29.0	33.3	44.7	98.93	94.7	25	107.51	111.93	81.52	43.00	65.36	50.81
9.	Backman	28.9	22.7	46.5	97.13	97.1	18	121.94	117.54	85.52	40.47	68.55	55.96
10.	Edison	26.3	24.2	48.3	98.20	98.6	30	105.30	111.59	72.80	41.19	61.64	51.58
11.	Emerson	27.2	30.3	35.1	97.24	109.6	15	114.27	129.36	88.98	47.77	75.48	61.14
12.	Hawthorn	24.9	18.8	29.3	97.88	105.8	16	109.93	120.41	89.99	41.85	64.14	61.15
13.	Nibley	23.3	19.8	36.5	97.28	117.6	13	106.76	118.84	84.65	41.37	64.29	50.15
14.	Riley	24.5	19.9	42.5	96.92	102.9	21	100.81	103.85	82.50	43.86	57.06	50.47
15.	Wasatch	19.2	18.8	27.4	98.00	110.8	9	128.59	136.49	97.73	48.38	79.54	67.54
	Total possible raw score points							147	158	125	64	100	96
	80th percentile							117.6	126.4	100.0	51.2	80.0	76.8
	20th percentile							29.4	31.6	25.0	12.8	20.0	19.2

linearly extrapolating the results to account for the fact that students had three additional weeks of exposure to the Title I program than they would have had if the test been administered exactly on the empirically established norm dates. Adjustments resulted in reducing the estimated impact (in NCE scores) of Model A by about 8%. Tests used in both the Model A and Model B comparison were taken from the Stanford Achievement Test (Madden, Gardner, Rudman, Karlson & Merwin, 1972) as shown in Table 3.

Table 3

Form and Level of the SAT Test Given for Model A and Model B Analyses

	Selection Test[a]	Pretest	Posttest
Grade 2	Primary I, Form A	Primary I, Form B	Primary I, Form A
Grade 3	Primary II, Form A	Primary II, Form B	Primary II, Form A
Grade 4	Primary III, Form A	Primary III, Form B	Primary III, Form A

[a]The test identified in this column refers to the selection test used under typical circumstances. As explained later in the report, a few children were legitimately selected for Model A using different measures. Selection test scores were unnecessary for Model B.

Analyses for the Model A evaluation were based on students selected in two ways. First, because the school district had been using their spring posttest as the selection test for next years students, they had traditionally only included those students in the analysis who had spring, fall and spring achievement test scores and remained in the same school from spring to spring.

Due to the substantial mobility within the district, this resulted in many children in spring included in Title I programs in the fall who were not tested during the spring testing, and other children who were tested in one Title I school in the spring and transferred to another Title I school for the next year. Although there is no guideline against limiting the analysis to those children who have spring selection test data and stay in the same school from one spring to the following spring, this procedure ignores the data of a substantial number of students for whom legitimate data is available.

The second selection method included additional students who had legitimate selection test scores even though they did not remain in the same Title I school from spring to spring. These additional students could be included from two groups. First, students who did not enter the Title I school until the fall, could still be included in the evauluation if they were selected based on an objective test that was separate from the pretest. Secondly, some students took the spring selection test in one of the district's Title I schools and then transferred out of that school into another Title I school in the district. In the past, these students who had transferred within the district had not been included in the analysis.

The selection of students to be considered in the Model B comparison schools could also be done in a number of ways which do not contradict the guidelines provided in Tallmadge and Wood (1976). Selection Method I recognized that even though comparison schools are reasonably similar, one comparison school could be somewhat higher on the average or could be distributed differently than another comparison school. Hence the lowest 15% in school A could have different scores from the lowest 15% in school B. In Selection Method I, children in all of the comparison schools were combined into one group and the percentage of children served in Title I schools served

by Title I programs during the peak enrollment period (January/February) was taken from the lower end of the test score distribution of the comparison school's fall test scores.

In Selection Method II, the number of students in each Title I school who were receiving Title I services during the peak enrollment period (January/February) was calculated as a percentage of that school's total student body. The median percentage of students being served in each of the eight Title I schools was taken as an average and this number was used in each of the comparison schools to select that percentage of students from the lower end of the distribution of the fall achievement test scores in each school. Since all of the comparison schools are essentially similar to each other and to the Title I schools (see Table 2), this method should provide a comparison group which is reasonably similar to the Title I group.

Selection Method III was similar with Method I in that children were selected from the lower end of the test score distribution after the comparison schools had been pooled into one group. However, this occurred in two stages. The same percentage of children was taken from the group of comparison schools that was taken from the Title I schools based on spring testing data. Natural attrition of students occurred between spring and fall in the comparison schools as it did in the Title I schools. A new group of children was selected from the lower end of fall test scores distribution in the comparison schools and added to the comparison group using the same percentage that was added to the Title I schools based on fall data.

Although Selection Method III for the comparison schools is clearly the most nearly like what happened in the Model A and consequently is best for comparing the results of Models A and B, it is not likely to be used if Model B were being implemented by the district. Selection Method I would be the

most plausible one for a district to implement. Moreover, Selection Method II, although not nearly as defensible empirically, is technically in agreement with the guidelines suggested by Tallmadge and Wood (1976). It is important to emphasize that Models A and B can be implemented correctly using very different groups of children as a basis for making the comparison about whether Title I programs are having any impact.

Results and Discussion

Shown in Table 4 are the NCE growth estimates for grades 2, 3, and 4 using Model A with only those children who were selected during spring testing and did not transfer to another school within the district (A), those children who were selected during spring testing plus those children who were legitimately selected during the fall or who were selected during the spring and then transferred to another Title I school within the district (A'), and Selection Method III in Model B which is probably the most rigorous and most similar to the way that children were selected for Model A. NCE growth estimates in both versions of the Model A results have been adjusted using a linear extrapolation to account for the fact that students were exposed to 33 weeks of instruction between the pre- and posttests rather than the 30 weeks of instruction that would have resulted had the test been given exactly on the empirically established norm dates. Dependent variables in all cases are subtests of the Stanford Achievement Test (1973 version). Average reading and math scores are unweighted arithmetic averages of the individual subtests which were administered at that grade level. Blanks in the table indicate that that particular subtest is not included in the level of test administered to that grade. For example, reading comprehension is not included in the level and form of the test administered to children in second grade.

Table 4

Title I Program Impact in NCE Gains for the Same Program Using Different TIERS Models and Selection Methods

	Grade 2			Grade 3			Grade 4		
	A	A'	B	A	A'	B	A	A'	B
Reading Part A	9.7	10.8	-1.8	-	-	-	-	-	-
Reading Part B	6.2	6.5	-8.4	-	-	-	-	-	-
Word Study	5.9	3.8	-.5	3.1	4.0	-3.5	-.4	-.5	-4.4
Reading Comprehension	-	-	-	2.7	3.5	-1.7	7.7	7.7	5.2
Average Reading	7.3 (n=50)	7.0 (n=76)	-3.6 (n=84)	2.9 (n=67)	3.8 (n=100)	-2.6 (n=103)	3.7 (n=115)	3.6 (n=146)	.4 (n=156)
Math Concepts	.7	2.6	-12.8[a]	5.2	5.6	1.0	4.4	4.0	-.6
Math Computation	3.5	6.7	-14.1[a]	7.5	7.1	-2.9	8.9	7.9	-3.6
Math Applications	-	-	-	4.1	2.3	-2.3	5.1	5.2	2.8
Average Math	2.1 (n=40)	4.7 (n=70)	-13.5[a] (n=75)	5.6 (n=63)	5.0 (n=89)	-1.4 (n=96)	6.1 (n=110)	5.7 (n=140)	-.5 (n=144)

NOTE: All numbers in parentheses refer to the number of students in Title I programs for whom data were available for that particular grade and evaluation model.

[a]Data for Model B on 2nd grade math should be viewed very skeptically because so few scores were available and pretest scores for available students were much lower than pretest scores in Title I schools. This was the only grade and test area where this occurred.

As can be seen from these results, Model A consistently yielded higher estimates of program impact than did Model B. Average differences for reading and/or math range from a low of 3.3 NCEs to a high of 10.9 NCEs (this is discounting the one difference of 15.6 NCEs on the average math for second grade students since scores for these subtests were based on very few control group students and had a number of anomalies that make the data questionable.

Using the results from Model B (which is theoretically the more rigorous model), it appears that the Title I program had no positive impact over and above what students would have achieved in the regular school program. Using the results of Model A, it appears that Title I is having a substantial positive impact.

Table 5 shows the estimated impacts of the Title I program using the three different selection methods for Model B described earlier. Depending on the selection method, very different children could be included in the comparison sample. Not only do the three methods differ in the children who are selected, but attrition in the three groups due to mobility or lack of test scores was probably systematically different in unknown ways from group to group so that the actual comparison group can be very different even though they are each selected in accordance with the guidelines.

The results in Table 5 show the average NCE gain on each subtest at each grade level using each of the three selection methods. Table 6 presents the same information in a different form. For each subtest at each grade level, the low estimate of Title I impact was set equal to 0 and the numbers for the other methods represent the difference between that method and the method having the low impact. Averages at the bottom of the table are an arithmetic average of each of the cell entries. The overall average to the right indicates the average across all cells for each method.

Table 5
Impact of Title I Program Using Three Different Selection Methods for Model B with Principal Axis Adjustment

SAT Subtest	Grade 2			Grade 3			Grade 4		
	I	II	III	I	II	III	I	II	III
Reading Part A	-2.1 (82)	.15 (82)	-1.76 (82)						
Reading Part B	-9.6 (82)	-6.6 (82)	-8.4 (82)						
Reading Word Study	-1.7 (84)	-1.2 (84)	-.47 (84)	-2.9 (103)	-2.4 (103)	-3.5 (103)	-5.1 (156)	-3.6 (156)	-4.4 (156)
Reading Comprehension				-1.1 (103)	-1.1 (103)	-1.7 (103)	.34 (156)	.37 (156)	5.2 (156)
Math Concepts	-14.8 (75)	-15.0 (75)	-12.8 (75)	-3.1 (96)	-.54 (90)	1.0 (90)	.7 (144)	-1.7 (144)	-.6 (144)
Math Computation	-15.4 (75)	-10.9 (75)	-14.1 (75)	-5.3 (95)	3.3 (95)	-2.9 (95)	-4.6 (143)	-2.4 (143)	-3.6 (143)
Math Applications				-7.0 (89)	-2.5 (89)	-2.3 (89)	3.0 (134)	2.0 (134)	2.8 (134)

Note: Selection Method I took median percentage of children served in each Title I school from the group of children in all comparison schools based on fall scores.

Selection Method II took median percentage of children served in each Title I school from each comparison school based on fall scores.

Selection Method III two-stage selection from group of children in comparison schools based on spring test scores and then adding some children based on fall test scores to make up for attrition over the summer.

Table 6
Differences in Model B Results Using Three Different Selection Methods

SAT Subtest	Grade 2 I	Grade 2 II	Grade 2 III	Grade 3 I	Grade 3 II	Grade 3 III	Grade 4 I	Grade 4 II	Grade 4 III
Reading Part A	0	2.3	.3	-	-	-	-	-	-
Reading Part B	0	3.0	1.2	-	-	-	-	-	-
Reading Word Study	0	.5	1.2	.6	1.1	0	0	1.5	.7
Reading Comprehension	-	-	-	.6	.6	0	0	.1	4.9
Math Concepts	.2	0	2.2	0	2.6	4.1	2.4	0	1.1
Math Computation	0	4.5	1.3	0	8.6	2.4	0	2.2	1.0
Math Applications	-	-	-	0	4.5	4.7	1.0	0	.8
	.04 4	2.06 1	1.58 0	.24 3	3.48 0	2.24 2	.68 3	.76 2	1.7 0

I = .32
II = 2.10
III = 1.84

Selection Method I took median percentage of children served in each Title I school from the group of children in all comparison schools based on fall scores.

Selection Method II took median percentage of children served in each Title I school from each comparison school based on fall scores.

Selection Method III - two-stage selection from group of children in comparison schools based on spring test scores and then adding some children based on fall test scores to make up for attrition over the summer.

Note: In each case, the lowest estimate for a particular subtest within each grade has been set equal to zero. Numbers in other boxes for that subtest and grade represent the difference expressed in NCEs between the low impact method and other methods.

73

As can be seen, Method II is consistently lower than either Methods I or III. Methods I and III yield similar results. The important fact is that although some systematic differences may persist over time using these three different selection methods, the attrition rates are influenced by so many other factors and directly affect estimates of impact that it is difficult to predict the differential impact of each method.

CHAPTER IV

DEGREE TO WHICH ASSUMPTIONS MADE BY MODEL A ARE MET IN UTAH TITLE I EVALUATIONS

Procedures

In addition to the comparison of Models A and B, 11 school districts were visited by project staff to investigate the degree to which districts utilizing Model A were violating assumptions of the model. During visits to these districts, staff members conducted structured interviews with district Title I directors, principals, and teachers in Title I schools to collect information about each of the Model A assumptions noted earlier. In addition, staff members observed Title I classrooms during the administration of the spring testing to determine the degree to which the procedures suggested in the publisher's manual were being followed during test administration and the degree to which teachers and students were on or off task during the test administration. Topics about which questions were asked during the interview with LEA staff members included:

1. The rationale for the particular test that was being used (both publisher and level);
2. Policy and practice for conducting make-up tests;
3. Policy and practice for checking the Title I data which was submitted for accuracy;
4. Adherence to the policy of not using the selection test as a pretest;
5. The time which tests were administered and whether or not adjustments were made when the testing date varied more than two weeks from the empirical norm date; and

6. The perceived relationship between the test content and the instructional emphasis in that school.

Research data was collected during on-site visits to 15 schools in 11 school districts. Eleven of the 13 districts required to report TIERS data to the State Office during the 1979-80 year were included. These 13 districts are a purposefully selected representative sample of all 40 districts in the state. Visits to all 13 districts were planned, but last-minute schedule changes with the district interfered and resulted in cancellation for 2 districts. The Title I director of each school district was notified by mail (Appendix #1) a month in advance of the on-site visit and familiarized with the purpose and methodology of the visit. Three weeks prior to the visit, the Title I directors were contacted by phone and given additional information as well as an opportunity to ask questions. Two weeks prior to the visit, the school principals were likewise informed (Appendix #2) and provided with a memo which they were asked to send to teachers and aides who would be visited (Appendix #3).

Each school district was visited for a day by one to three data collectors. The data collectors arrived at the schools at 8:00 a.m. and individually interviewed the principals and selected teachers and aides for approximately one hour. The data collectors then visited prearranged classrooms and unobtrusively observed the students' and teachers' testing behavior for approximately 45 minutes. After a short break, the data collectors administered a Format Familiarity Test to several students in an empty room for approximately 20 minutes (see Chapter V for a more complete description of this component of the project). The Title I teachers and aides were then interviewed again for approximately 30 minutes and asked to fill in a

curriculum content survey. At the end of the day, the Title I director was interviewed at the district office and informed of the day's events.

Five types of data were collected during the visit.

1. Open-ended Interview

The purpose of the open-ended interview was to: (a) determine the awareness of Title I personnel to possible Title I Model A violations, (b) examine the roles of specific personnel in each district in collecting, collating and distributing TIERS data, (c) investigate the coherence and communication among the Title I personnel within a district, (d) develop an impression of the degree of involvement and satisfaction of the Title I personnel with the program and evaluation techniques, and (e) identify common problems, complaints and sources of possible future difficulties (see Appendix 4 for the interview guide sheet used).

2. Determination of On Task Behavior During Testing

Title I testing periods were observed by one to three trained observers in 20-minute blocks. Five students and one teacher were observed in each block. On and off task behavior was recorded in 5-second intervals, using a tape recorder and earphones as a pacer. The observers were trained at the university prior to the on-site visit using both video tapes and actual classroom observation and attained a minimum of .85 interrater reliability. The data collection form and definitions for student and teacher on task and off task behavior are included in Appendix 5. Further information on the procedures used for collecting these data is contained in Chapter VI.

3. Quality of Test Administration Checklist

Environmental, instructional, and situational variables which contribute to high-quality standardized test administration were recorded using a dichotomous checklist. Variables included in the checklist were identified based on standardized test administration manuals and textbooks on standardized test administration. Data collection procedures practiced at the univerity prior to the on-site visit attained a high degree of interrater reliability (.90 or better). The checklist used is included in Appendix 6.

4. Format Familiarity Test (FFT)

The FFT was administered following the Title I testing period. The Title I teacher or aide was asked to supply four to six average Title I students so that they could be administered a test to determine if a student demonstrated knowledge of reading phonics differed depending on the format in which the test was administered. After a short rest period, these students were tested in an empty room for approximately 15 minutes. Additional information about the procedures and results of this activity are included in Chapter V.

5. Curriculum Content Interview

The Title I teachers and aides were queried about the content of their reading curriculum and the relative emphasis and importance they accorded the various content areas. The interview format used during this discussion is shown in Appendix 7.

In addition to these data collection activities, a variety of additional assistance was provided to the Salt Lake City School District in implementing TIERS. One of the major activities was assisting in developing a workable

procedure and set of definitions for computing student/teacher ratios in Title I programs. Based on the resulting procedures (see Appendix 8), Title I teacher leaders were interviewed and data was collected for computing the student/teacher ratios of their Title I programs.

The districts and schools visited, the dates of visits, and the personnel interviewed to collect data in the five areas listed above during this component of the project are listed in Table 7. The type and amount of data collected from each school and grade level are listed in Table 8.

Results

Interviews with LEA Staff

The open-ended interviews consisted of five parts: (1) students' reaction to testing, (2) district personnel reaction to testing, (3) selection of students, (4) test administration, and (5) procedures for submitting TIERS data. Data was collected from three categories of personnel: (1) Title I directors, (2) principals, and (3) Title I teachers and aides. The results were as follows.

Student reaction to testing. The majority (55%) of the Title I directors were not familiar with their students' reactions to testing. Of the remainder, 50% said the students were not negative toward testing and understood it, while 100% felt the students generally behaved well and tried their best.

The principals were generally positive about the students' reaction to the testing (62%) and their understanding of the purpose of the test (62%). All of the principals felt the students generally behaved well, but only 79% felt the students tried their best.

Table 7

Title I Field Visit Information

District	Date Visited	Elementary School Visited	Personnel Interviewed
Alpine	4/29	Greenwood Greenwood	Dr. Tregoskis - Title I Director Mrs. Brannon - Title I Teacher Mr. Crandal - Principal
Box Elder	4/28	Lincoln Lincoln	Mr. Harding - Title I Director Mr. Stanger - Principal Mrs. Hogart - Title I Teacher
Duchesne	4/2	Duchesne Duchesne Duchesne Myton Myton	Mr. Hansen - Title I Director Mrs. Geiest - Principal Mrs. Meldrum - Title I Teacher Mr. Duke - Principal Mrs. Roberts - Full Time Aide
Morgan	4/24	Morgan Morgan	Mr. Jeffrey - Title I Director Mr. Warna - Principal Mrs. Adamson - Title I Teacher
Murray	4/9	Viewmont Viewmont Viewmont	Dr. Bertleson - Title I Director Mr. Campbell - Principal Mrs. Middleman - Title I Teacher Mrs. Froelich - Title I Teacher
Park City	5/12	Old School Old School	Dr. Falls - Title I Director Dr. Falls - Principal Mrs. Shonon - Title I Teacher
Provo	4/17	Timpanogas Timpanogas	Dr. Brimley - Title I Director Mr. Gunther - Principal Mrs. Murdoch - Title I Teacher
Salt Lake City	5/6	Parkview Parkview Parkview Washington Washington Bennion	Mrs. McDonald - Title I Director Mrs. Weggeland - Principal Mrs. Crawford - Title I Teacher Mrs. Taylor - Title I Teacher Dr. Comb - Principal Mrs. Jackson - Title I Teacher Mrs. Dolee - Title I Teacher
S. Sanpete	5/6		Dr. Graham - Title I Director Dr. Graham - Principal Mrs. Richardson - Title I Teacher
S. Summit	5/5	S. Summit S. Summit	Mrs. Link - Title I Director Mrs. Walker - Principal Mrs. Marchant - Full Time Aide
Wasatch	5/7	North North	Mrs. Baird - Title I Director Mr. Dayton - Principal Mrs. Ivy - Title I Aide

Table 8

Data Collected Regarding TIERS Implementation in Eleven Utah School Districts

	PERSONEL INTERVIEWED				# OF STUDENTS ADMINISTERED FFT		# OF ON/OFF TASK PERIODS						CURRICULUM ESTIMATE			TEACHER CHECKLIST		
	Title I Director	Principal	Title I Teacher(s)	Title I Aide(s)	2	3	1	2	3	4	5	6	2	3	4	Session 1	Session 2	Session 3
ALPINE	X	X	X		6			2					X			X	X	
BOX ELDER	X	X	X		4	4		1	1				X	X		X	X	
DUCHESNE	X	X	X	X				1	1		2	2		X		X	X	X
MORGAN	X	X	X				2						X			X		
MURRAY	X	X	X	X				1		1			X		X	X	X	
PARK CITY	X	X		X		6		1								X		
PROVO	X	X	X					2	2				X	X		X	X	
SALT LAKE	X	X	X					2	2	2			X		X	X	X	X
S. SANPETE	X	X	X		8			1	1				X	X		X	X	
S. SUMMIT	X	X		X	6			1					X	X		X		
WASATCH	X	X		X	6	6		2	1							X	X	

Only 22% of the teachers felt the students reacted positively to the testing, while 79% felt the students understood the purpose of the test and 88% felt they generally behaved well. Only 44% felt their students tried their best.

District personnel reactions to testing. All of the Title I directors felt the testing was worthwhile although some said there was too much testing. Only 13% felt that the test results were used for anything other than to compare gains. Thirty-seven percent of the directors did not know if test results were discussed with parents, but all of the other directors responded that test scores were individually discussed with the students and/or parents.

Of the principals, 86% felt the testing was worthwhile. Only 33% said they used the test results in any way other than to compare gains, while 83% said the results were individually discussed with either the students or parents.

The teachers definitely had the most negative reaction to the testing. Only 38% of the teachers felt that the testing was worthwhile, only 29% said they made any special use of the test scores, and only 44% said they actually discussed the test scores with the parents or students. Many said they would, however, if the parents ever bothered to visit or inquire.

Selection of students for Title I programs. The Title I directors were the best informed about the process involved in selecting the students. This generally involved a screening test and a teacher evaluation. The directors, principals, and teachers unanimously felt that the selection process resulted in selecting the right students.

Only one district reported not keeping the selection and pretest different because they did not feel this was necessary. The directors, principals and teachers could only make guesses ranging from 0% to 50% as to the percentage of students that were new move-ins, testing out of level, and/or transfers. Interestingly, no one knew if or how many students were tested out of level.

The biggest discrepancy between the three personnel levels concerned how new move-ins were selected. The answers differed more qualitatively than quantitatively. The directors responded with the proper technique, i.e., each new move-in was tested on one or more selection measures and evaluated by the teacher; then the results were weighted and it was determined whether the child met the established criteria. The principals responded with what they hoped practically happened. Each new child was given a test which determined whether they would be in Title I. The teachers explained what really happened. Sometimes, the new students were given a test, but generally after two weeks the teachers themselves knew whether or not the student should be in Title I and the test was frequently not given.

Test administration. Title I directors unanimously agreed that tests were given on the empirical norm dates, that students who were absent during testing were tested immediately upon their return and that the students always got make-ups. Only one district extrapolated test scores and most of the rest were not certain as to when or why this was necessary. Again, there was confusion as to what, how often, and why they gave out-of-level testing.

Every district gave a different reason for selecting their particular type, form, and level of test. Some of the reasons were: easiest to administer, familiarity, one that allows half-year testing, convenience and

Table 9

Percent of Agreement by Title I Directors, Title I Teachers and Aides, and Principals to Open-Ended Interview Questions Which Could Be Answered Yes or No

Question	RESPONDENTS		
	Directors (n=11)	Principals (n=13)	Teachers & Aides (n=16)
1) Do students feel positive about testing?	50 (25)	62 (0)	22 (10)
2) Do students understand the purpose of the testing?	50 (50)	62 (12)	75 (0)
3) Do students behave well during testing?	100 (37)	100 (37)	88 (0)
4) Do students try their best during testing?	55 (50)	29 (12)	44 (0)
5) Is achievement testing worthwhile?	100 (0)	86 (12)	38 (0)
6) Are test scores used for purposes other than compliance with TIERS?	13 (0)	33 (12)	29 (0)
7) Are test results discussed with the students and/or parents?	63 (0)	83 (0)	44 (0)
8) Are tests administered within guidelines for empirical norm dates?	83 (12)	75 (50)	100 (75)
9) Are test scores extrapolated where necessary?	100 (87)	100 (87)	0 (87)
10) Are students given adequate preparation for testing?	71 (0)	50 (25)	50 (25)
11) Are teachers given any special preparation for test administration?	71 (0)	33 (0)	29 (0)
12) Are forms and procedures for submitting data to SEA good?	80 (12)	100 (87)	100 (75)
13) Are students who miss testing, tested upon their return?	100 (50)	100 (12)	100 (12)
14) Does the selection process result in appropriate students being selected?	100 (50)	100 (0)	83 (25)
15) Are selection and pretest separate?	75 (25)	100 (50)	66 (62)

NOTE: The number in each cell represents the percentage of people who responded to the question who answered it affirmatively. The numbers in parentheses indicate the percent of people who felt they did not have enough information or knowledge to answer.

funds, school curriculum match, Northwest Lab's recommendation, and literature review.

Seventy-one percent of the Title I directors thought that special effort was made to prepare the students or teachers for testing while only 50% of the principals and teachers felt that the students received little or no preparation for testing. Furthermore, only 25% of the principals and teachers felt the teachers were given any special preparation for the testing.

Procedures for submitting data to SEA. In all instances the Title I directors are responsible for submitting data to the USOE, and this year's reporting format is thought to be far better than that of the previous year. The accuracy of the testing data is supposedly ensured by random error checks in some districts, while in most it is trusted to the teachers and only glaring errors are noticed. None of the districts had any evidence that random error checks were actually done.

Summary of open-ended interviews. The results from the open-ended interviews are summarized in Tables 9 and 10. Table 9 shows the percentage of people in each category who felt they had enough information to respond who answered each question affirmatively (the numbers in parentheses show the percentage of people who said they did not know enough to answer). Many of these questions are directly related to the validity of standardized testing in general and specifically the implementation of the TIERS models.

Although these data indicate that there may be some minor to moderate violations of TIERS assumptions regarding Model A, the lack of agreement among LEA personnel on many of the questions is also disturbing. In addition to questions which could be answered "yes" or "no" a number of other questions were asked regarding procedures and opinions regarding TIERS implementation. The most disturbing fact about this information was the lack of agreement among various LEA staff members about what was really taking place. Table 10

Table 10

Agreement Among School Personnel On
Open Ended Interview Questions

Question	D/P*		D/T*		P/T*	
	A**	DK**	A**	DK**	A**	DK**
1) How are the students selected for Title I programs?	42	12	25	75	66	25
2) What percent of students are tested out of level?	0	75	0	75	100	87
3) How are new move-ins selected?	25	12	40	50	63	62
4) What is the percentage of new move-ins in the district?	66	62	0	75	50	50
5) What is the percentage of turnover in Title I programs?	0	87	—	100	0	87
6) What percent of the students new get a make-up test?	—	100	—	100	—	100
7) Who is responsible for turning data into USDE?	100	75	0	87	0	75
8) How did you select your particular achievement test?	100	87	—	100	—	100
9) What percent of students miss testing?	—	100	—	100	—	100
10) How is accuracy of data checked?	100	87	0	87	80	75
AVERAGE AGREEMENT	54%		9%		40%	

*D/P = Percent agreement between Directors and Principals
*D/T = Percent agreement between Directors and Teachers
*P/T = Percent agreement between Principals and Teachers

**A = Agree
**DK = Don't know or not obtained

shows average agreement between directors and principals, directors and teachers, and principals and teachers. Numbers in each cell represent the percentage who were in agreement, while the numbers in parentheses indicate the percentage of time one or both of the pair did not know enough to respond. The low levels of agreement among LEA staff on these items make definitive conclusions about what is really happening difficult. The preceding narrative about each of the areas discussed in the interviews represents the project staff's best estimate and indicates some moderate problems.

The low levels of knowledge of agreement among LEA staff raises additional questions. For example, only 13% knew if test scores were extrapolated (question #9, Table 9), and only 21% knew how many students were tested out of level (question #2, Table 10). Only 42% of the school personnel agreed as to how new move-ins were selected (question #3, Table 10). Overall, Title I directors and principals had the greatest agreement about the testing procedures (54%--perhaps, because they are more familiar with what the procedures should be and thus answered "correctly"). The Title I directors and teachers were least in agreement (9%--perhaps because the teachers responded with what actually happened at the school).

Determination of On-Task Behavior During Testing

Table 11 shows the average percent of time that teachers and students were on task during test administration. "On task" was defined for both teachers and students according to discrete, observable behavior. For example, movement for a teacher was "on task" if she was standing in front of the room or pointing at nonattenders or providing a pencil, but was "off task" if she was standing with her back to the students or where students' faces couldn't be seen, or sitting down at her desk correcting papers during the

Table 11
Percent of Time Students and Teachers Are
On Task During Test Administration -

	Students	Teachers
Total	76.5 (n = 32)	52.5 (n = 12)
Grades 1 and 2	75.3 (n = 15)	70.0 (n = 5)
Grades 3 and 4	75.3 (n = 12)	41.7 (n = 5)
Grades 5 and 6	82.7 (n = 5)	35.5 (n = 2)

Table 12

Percent of Time Students and Teachers
are Having Contact During
Test Administration

	X Based on Student Observation	X Based on Teacher Observation
Total (n = 32)	3.5	5.9
1st & 2nd Grade (n = 15)	5.1	8.9
3rd & 4th Grade (n = 12)	1.7	5.7
5th & 6th Grade (n = 5)	2.9	0

test. On-task and off-task behaviors were defined based on suggestions in publisher's test administration manuals and textbooks on standardized group achievement test administration.

Students were on task about 75% of the time and teachers were on task only about 52% of the time. The percent of on-task behavior increases for students in the higher grades slightly and decreases for teachers in the higher grades.

Quality of Test Administration

Table 12 shows the mean percentage of time which students and teachers have contact during the standardized test administration. The discrepancy between the estimate based on student observation and teacher observation is because only five Title I students out of the total class were observed but during the teacher observation, contact with any student was counted. Table 13 shows the results of a dichotomous designation of various activities which teachers should be doing before and during test administration. As can be seen, many of these items are done infrequently with some of the most important items (e.g., under student preparation see #4 . . . "Explain the reason for the test") being done least frequently. Items on the checklist were taken from publisher's recommendations and standard textbooks on test administration.

Match Between Curriculum Emphasis and Testing

Table 14 shows the percent of time Title I teachers report they spend in class teaching five areas of reading skills and the importance they place on these skills. Table 15 shows the emphasis that the reading subtests of six standardized achievement tests place on the same five skill areas as determined by the percent of questions directed toward sampling a student's skill in each of the areas. Table 16 shows the differences in emphasis on

Table 13

Degree to Which Recommended Practices in Standardized Test Administration Are Followed During Testing

DID THE TEACHER DO THESE BEFORE ADMINISTERING THE TEST?

n = 38

Class Environment		
68%	1.	Arrange the students' desks so they are not touching.
58%	2.	Position the desks to face the same direction (every booklet and student's face can be seen from the front of the room).
78%	3.	Assure that the room is comfortable (temperature, light, noise).
56%	4.	Post a "Testing, Do Not Disturb" sign on door.
84%	5.	Have a visible supply of pencils.
94%	6.	Have a visible clock or watch with a minute hand.
89%	7.	Create a generally positive climate that promotes good work habits and is without pressure or tension.
40%	8.	Seat the most frequently nonattending students in the front.
Student Preparation		
45%	1.	Provide an opportunity for using the bathroom, drinking water, and sharpening pencil.
97%	2.	Provide all students with a pencil and an eraser.
59%	3.	Ask students to remove nontesting material from desks if appropriate.
41%	4.	Explain the reason for the test (to use the information to help teach students).
71%	5.	Obtain the attention of the entire class for 1 minute prior to directions (all students watching teacher).
89%	6	Pass out test booklets in less than 2 minutes and in an orderly and efficient manner.
49%	7.	Verbally reward attentive behavior.
Reminders		
40%	1.	Not to leave their seats but to raise a hand if something is needed.
47%	2.	What to do if they finish before time is up (TT only).
27%	3.	To check their work if they finish before the time is up (to see if every question is answered only once).
10%	4	That some of the items will be more difficult than their daily work.
10%	5.	To skip an item that they don't know and go on to the next one.

Table 13 (continued)

Did the teacher do these WHILE administering the test?

Positive Atmosphere

62% 1. Praise individual students for appropriate behavior.
53% 2. Praise class for listening and working.
63% 3. Smile frequently.
89% 4. Make less than two reprimands, threats, or criticisms during the subtest.
95% 5. Speak with a gentle, but firm voice.
75% 6. Use physical touch to prompt and reward on task behavior.
78% 7. Start the test directions within several minutes of sitting down so that students did not become restless with preparation activities.
64% 8. Quickly supply a student with pencil or eraser when needed.
58% 9. Stand near front of room where all students can see easily.

Reading Directions

10% 1. Look at class between sentences.
89% 2. Survey the class to check if directions were followed (i.e. "Put your finger [illegible]sample," "fill in the circle," "write your name," "turn to page 12").
23% 3. [illegible] aide to nonattenders.
79% 4. [illegible]o next direction only after all students are ready.
73% 5. [illegible]nt printed directions with verbal and visual explanations when students do not understand the procedure.
41% 6. Change wording of directions to a vocabulary the students are familiar with (i.e. "circle" instead of "oval" or "box" instead of "frame".

Reading Test Items

74% 1. Look up after each question and glance around room.
54% 2. Follow the exact wording of questions as stated in the manual (Never define or explain words or illustrate procedures.)
67% 3. Allow approximately 10 seconds between items.
50% 4 Never repeat a question unless the directions specify to do so.
33% 5. Alert aide to nonattenders or to students with raised hands.

Timed Tests

62% 1. Set clock for correct time requirement.
73% 2. Watch students during entire test to detect speeding, slow answering, day dreaming and cheating.
50% 3 Alert aide to nonattenders or to students with raised hands.

End of Test

53% 1. Praise students for working hard.
84% 2. Collect booklets in a directed manner.
56% 3. Provide a directed, stand-up, rest period.

Table 14

TEACHER RATINGS OF THE IMPORTANCE OF VARIOUS CONTENT AREAS ASSESSED BY STANDARDIZED READING TESTS

District (N)	Test Used	% Time Spent on Each Area					Importance of Each Area (1 = high; 5 = low)				
		Phonics	Vocabulary	Literal Comprehension	Inferential Comprehension	Structural Analysis	Phonics	Vocabulary	Literal Comprehension	Inferential Comprehension	Structural Analysis
1 - S. Summit (2) (2nd & 3rd grade)	ITBS	25	30	13	13	18	1	2	4	5	3
2 - S. Sanpete (1) (2nd grade)	G-M	10	10	35	35	10	2	1	4	5	3
3 - S. Sanpete (1) (3rd grade)	G-M	5	10	35	40	10	5	3	2	1	4
4 - Provo (1) (2nd & 3rd grade)	SAT	40	20	20	0	20	1	2	3	5	4
5 - Morgan (1) (1st grade)	G-M	60	20	7	3	10	1	2	4	5	3
6 - Duchesne (2) (3rd & 4th grade)	CAT	30	30	13	7	20	1	2	3	5	4
7 - Box Elder (2) (2nd & 3rd grade)	ITBS	35	25	25	5	10	1	2	3	5	4
8 - Alpine (1) (2nd grade)	SAT	15	30	10	5	40	1	3	4	5	2
9 - Murray (1) (2nd grade)	Woodcock	90	10	0	0	0	1	2	3	5	4
10 - Murray (1) (4th grade)	Woodcock	50	50	0	0	0	1	2	4	5	3
11 - Salt Lake City (3) (2nd grade)	SAT	37	20	17	8	18	2	2	4	4	4
12 - Salt Lake City (2) (3rd grade)	SAT	20	25	20	10	25	4	2	2	4	2
13 - Salt Lake City (3) (4th grade)	SAT	15	16	30	20	16	5	3	3	2	4
TOTAL		33	23	17	10	15	2.0	2.2	3.3	4.4	3.4

Table 15

Percent of Items Devoted to Various Areas of Reading for Six Standardized Tests

2nd GRADE

	CAT	SAT	WOODCOCK	ITBS	GATES-M	SRA
Vocabulary	21	20	20	20	80	28
Comprehension (both literal and inferential)	29	47	24	46	0	28
Phonics	35	33	56	34	20	22
Structural Analysis	15	0	0	0	0	0

3rd GRADE

	CAT	SAT	WOODCOCK	ITBS	GATES-M	SRA
Vocabulary	21	19	20	41	51	54
Comprehension (both literal and inferential)	37	48	24	59	49	46
Phonics	27	33	56	0	0	0
Structural Analysis	15	0	0	0	0	0

Table 16

Differences in Emphasis of Skills Taught by Teachers and Sampled by Standardized Tests for Grades 2 and 3

District	Differences in Percent of Time Devoted[a]				Average[c] Descrepancy	Differences in Importance[b]			
	Phonics	Vocab.	Compre.	S.A.*		Phonics	Vocab.	Compre.	S.A.
S. Summit (2nd)	-9	+10	-20	+18	14.3	+1	+1	-3	+1
S. Summit (3rd)	+25	-11	-33	+18	21.8	+3	0	-3	+1
S. Sanpete (2nd)	-10	-70	+70	+10	40.0	0	0	0	+1
S. Sanpete (3rd)	+5	-41	+26	+10	20.5	0	-1	+1	0
Provo (2nd)	+7	0	-27	+20	13.5	+1	+1	-2	0
Provo (3rd)	+7	+1	-28	+20	14.0	+1	+1	-2	0
Duchesne (3rd)	+3	+9	-17	+5	8.5	0	+1	-1	0
Box Elder (2nd)	+1	+5	-16	+10	8.0	+1	+1	-2	0
Box Elder (3rd)	+35	-16	-29	+10	22.5	+3	0	-2	0
Alpine (2nd)	-18	+10	-32	+40	25.5	+1	0	-3	+2
Murray (2nd)	+34	-10	-24	0	17.0	0	+1	-1	0
Salt Lake City (2nd)	+4	0	-22	+18	11.0	0	+1	-3	0
Salt Lake City (3rd)	-13	+6	-18	+25	15.5	-2	+1	-1	0

*Structural Analyses

[a] The percent difference scores are computed by subtracting the percentage of test items devoted to sampling a reading skill area (Table 15) from the percent of time the district spends teaching those skills (Table 14). Hence, a positive score means that a district is placing greater emphasis on teaching a given skill area than is sampled by the test indicated by the number of items on the test pertaining to that skill area.

[b] The differences of importance scores are computed by subtracting the rank ordering of perceived importance of skill areas by Title I teachers (Table 14) from the rank ordering of importance of the skill areas in the standardized tests as determined by the number of test items sampling each skill area. Hence, a positive score means the district thinks a certain skill area is more important than indicated by the test.

[c] The average difference between instructional emphasis and test emphasis was not used to compare districts. The greater number in this column the more descrepant is the district's instructional emphasis from the emphasis suggested by the number of test items.

skills taught by teachers and sampled by standardized tests. Although the techniques used to collect these data are admittedly somewhat crude and based on limited numbers of teachers in each district, nonetheless, the results suggest that some schools are using tests which are completely inappropriate. For example, the third grade teacher in South Sanpete spends 75% of her time on comprehension and ranks it as the most important of the reading skills. Yet, the test used by the district does not include any items designed to measure reading comprehension.

Discussion

The results of the data collected on the field visits indicate that districts can do much to improve the preparation for and implementation of standardized achievement tests used in conjunction with TIERS. Of greatest concern is the apparent lack of awareness of proper testing procedures among school personnel. Even in those districts in which programs appear to be properly implemented, the lack of agreement and or lack of knowledge among personnel as to what the program is actually doing raises substantial concern.

The second area of concern is the appropriateness of the standardized achievement tests selected by each school district. Only one school district chose a test because of data suggesting its suitability. Most districts selected tests because of price, convenience, or habit. If all achievement tests were equally suitable, these would be excellent determinants of test selection. However, there is strong evidence that the differences in content sampled by the tests and the differences in the format in which the tests are administered (see Chapter V) produce significant differences in scores that cannot be accounted for by normative scaling. In some school districts and

grades, there is a wide diversity between what is taught in the classroom and what is sampled by the achievement test used in that district and grade.

A final concern is the performance of the individual teachers before, during and after testing. In many districts opportunities were not afforded for the teachers to be adequately prepared or familiarized with the test and testing procedures. The data collected during the visits to LEAs suggest that standardized testing procedures are frequently violated and both teachers and students demonstrate fairly high levels of off task behavior. In summary, awareness of TIERS guidelines needs to be improved, better communication is needed between Title I directors, Title I teachers and principals, better criteria for selecting tests should be adopted, and efforts are needed to assure that standardized test administration procedures are used.

CHAPTER V

THE EFFECT OF ITEM FORMAT ON STUDENTS' STANDARDIZED READING ACHIEVEMENT TEST SCORES

Cronbach (1971) defined the universe of behavior sampled by a test as including all abilities required to perceive the stimulus items and formulate responses. For an achievement test to be a valid indicator of what a student knows, it must be assured that the student gets answers correct or incorrect because of what he or she knows. If variables besides the student's knowledge content affect the student's score, correct interpretation of the results becomes more difficult, and one must question the validity of the test for making decisions that depend on the student's knowledge (e.g., educational placement and instructional programming decisions).

During the preparation for the on-site visits to various districts described in Chapter IV, it became clear that the format in which a standardized test question was asked, was one such factor that might influence test scores and confound what a student appears to know. Consequently, a component was added to the work scope of the project to investigate the effect of item format on students' standardized reading achievement test scores. This chapter describes the rationale for the study, the procedures and results, and the implications for the TIERS. In addition to the original study, the results of a follow-up replication of the study with Title I students in Texas are included. Although the follow-up study was not conducted with project funds, the procedures for the study were developed during the project and the results of the second study underscore the importance of format differences in interpreting test scores.

Previous Research

Previous research has investigated a number of factors which may affect how well a student does on standardized achievement tests besides what the student knows. The research summarized briefly below considered variables inherent in the tests themselves and the way questions were asked as opposed to student characteristics (e.g., locus of control, motivation, anxiety) or setting variables (e.g., time of day, student/examiner ethnicity match).

Order of Test Items

Whether test items are presented from easiest to most difficult, most difficult to easiest, or in scrambled order may affect the students' test scores. Barciskowski and Olsen (1975) reported that students feel tests are harder when the test items are presented in a decreasing order of difficulty. Towle and Merrill (1975) showed that the order of presentation can affect the anxiety level of the test and thus influence performance. Some researchers have reported that increasing item difficulty elicits better performance (Holliday & Partridge, 1979), while others have found no effect at all (Gerow, 1980; Dambrot, 1980).

Test Item Format

Millman and Setijadi (1966) showed that Indonesian students perform better than American students on open-ended math problems while American students perform better on multiple-choice problems. Evidently, the format in which questions were asked was a determinant of students' scores.

In some instances the format of a test item decreases test scores by demanding additional intellectual skills to answer the question. Poage and Poage (1977) found that math questions containing pictures were missed more

often than comparable questions without pictures on the Michigan State Assessment Test. They hypothesized that items with pictures resulted in lower scores because they required cognitive capacities involving the use of perspective that were not fully developed in younger children. In a related study, Kierscht and Vietze (1975) investigated whether younger children perform better on questions that use three dimensional objects as the test stimulus rather than two dimensional pictures. Kierscht and Vietze (1975) found that preschool children performed better on the Slossen Intelligence Test which uses objects, than on the Peabody Picture Vocabulary Test which uses pictures.

Research has also demonstrated that students perform better on multiple choice than on short answer type question format covering the same content (Kumar, Rabinsky, & Pandley, 1979; Loftus, 1971). Estes and DePolito (1967) suggested that such differences occurred because the short answer format required the student to recall the information from memory, whereas the multiple choice format only required the student to recognize the correct answer. Kintsch (1970) similarly suggested that recall items required the student to both search for and retrieve information, whereas recognition items only required the student to discriminate between the presented information. Halpin and Halpin (1979) demonstrated that students performed best when the type of test content was matched with the type of item format (i.e., when recognition item formats sampled concept level questions; and recall item formats sampled knowledge level questions).

Frisbie (1973) hypothesized that "multiple choice questions limit the universe of comparisons" because they only require an individual to recognize the correct response. When answering a true-false question, however, an individual must eliminate the incorrect response by forming a counter example

from a wide universe of possibilities. Benson and Crocker (1979) administered multiple-choice, true-false and matching type questions of identical content to students and also found a statistically significant difference in the scores for each item format. Students scored highest on the matching questions and worst on true-false questions.

Question and Answer Form

All test items require that a question be asked and an answer be given. The way in which questions are asked or answers given vary from test to test. For example, Table 17 displays four different question and answer formats from the word analysis subtests of three second grade standardized achievement tests. Differences in the form for either questions or answers may affect the scores students receive on standardized tests.

Johnson, Pittleman, Ackwenker and Perry (1978) administered the vocabulary subtest of the Metropolitan Achievement test with three different types of question forms. The question forms were synonym, synonym in context, and cloze. The answer form in all cases was multiple choice. The students performed best on the synonym format in which they were presented a stimulus word and asked to find the response alternative which was closest in meaning to the word. The synonym in context format was of intermediate difficulty. In this form, the stimulus word was imbedded in a sentence. Students scored lowest on the cloze format, which is a "fill in the blank format". In a nonempirical article, Roid and Hadaynna (1978) concluded that multiple-choice item stems sampled content best when the item stems had fixed syntactical structure (as found in synonym in context formats) and when adjectives and verbs were used rather than other parts of speech as the target words in cloze tests.

Table 17

Sample Test Item Formats Taken From the Phonics Analysis Subtests for Three Standardized Achievement Tests

---

FORMAT #1: Stanford Achievement Test, Level II, Form A.

"Mark under the word CAT ."

CAR	CAN	CAT
o	o	o

FORMAT #2: California Achievement Test, Level 12, Form C

"Find the word with the same beginning sound as the word CAT ."

CAN	TAN	TACK
o	o	o

FORMAT #3: Stanford Achivement Test, Level II, Form A.

"Mark the space under the word that has the same sound as the underlined sound." CAT

CAR	SAT	NAP
o	o	o

FORMAT #4: Woodcock Reading Mastery Test

"...does this word sound." CAS

Changes in the answer form may also affect the child's score on an achievement test. Grier (1975), Weber (1977) and Catts (1978) found that additional distractors decreased what the child appeared to know. Williams, Davis, Anderson and Favor (1978) and Lai-Min and Coffman (1974) administered the original and a modified version of the Iowa Test of Basic Skills to elementary school students. Williams, et al. found that an additional "all incorrect" distractor decreased scores, and suggested this decrease was attributable to students' unfamiliarity with that format. Lai-Min and Coffman added an "I don't know" distractor and also found that scores decreased. Weiten (1979) administered multiple-choice questions with either compound or single response forms and found that students scored lower on the compound response form.

Summary

Achievement tests are usually interpreted as an indication of what children know and are frequently used to make important decisions concerning the placement, advancement, and educational programming of children. If, as previous research has indicated, types of test item formats influence what a child appears to know, the types of formats used by an achievement test should be considered in selecting and interpreting achievement test scores.

Based on previous research, it appears that how the test items are organized, the way in which test items are asked, and the method by which they are answered all affect students' performance on tests. The research described herein expands the findings of previous research by comparing the effects on students' scores of four test item formats taken from three commonly used standardized reading achievement tests.

Method

Test Construction

The Stanford Achievement Test (level II form A), the California Achievement Test (level 12 Form C), and the Woodcock Reading Mastery Test were analyzed. Four different types of question formats for testing knowledge of phonic sounds were identified (see Table 17). The content (knowledge of the phonic sounds) was classified into six categories: consonant sounds, vowel sounds, consonant digraphs, vowel digraphs and diphthongs, controlled vowels, and variant vowels.

A content bit was selected from each of the six categories. A Format Familiary Test (FFT) was constructed so that each of these six content bits were tested using each of the four formats. In other words, identical content was tested for each child using four different formats. For instance, the students' mastery of the consonant digraph "th" was tested using all four format types. There were a total of 24 separate questions as shown in Table 18. Format #4 was scored in two ways resulting in five scores for each phonic sound. First, format 4A was scored correct if the student pronounced the target sound correctly. Second, format 4B was scored correct only if the student pronounced the entire word correct (this is the scoring procedure suggested by the manual).

Procedures

The Format Familiary Test was administered to two groups of students. In Study I, the Format Test was administered to 37 second grade Title I students in May, 1980. The students were from eight school districts in Utah. They were primarily Caucasian and evenly split between rural and urban communities. In Study II, the Format Test was administered to 31 second grade Mexican-

1. Stanford Achievement Test Level II
   "Mark under the word _____."

2. California Achievement Test Level 12
   "Find the word with the same beginning sound as the word ____."

3. Stanford Achievement Test Level II
   "Mark the space under the word that has the same sound as the underlined sound."

4. Woodcock Reading Mastery Test
   "How does this word sound?"

Table 18

Item Matrix of the Six Phonic Sounds

Tested with Four Format Types

CONSONANT SOUND	SHORT VOWEL SOUND	CONSONANT DIGRAPH	VOWEL DIGRAPH AND DIPHTHONG	CONTROLLED VOWELS	VARIANT VOWELS
sn	u	th	ou	ar	au
sow 0 snow 0 show 0	rug 0 rig 0 rag 0	torn 0 horn 0 thorn 0	sport 0 spout 0 spot 0	bran 0 born 0 barn 0	laugh 0 lag 0 leaf 0
shake 0 state 0 snack 0	bad 0 big 0 nut 0	thick 0 tow 0 torn 0	loud 0 putt 0 pot 0	for 0 from 0 mark 0	told 0 caught 0 talk 0
snip ship 0 snack 0 sip 0	luck rut 0 lick 0 mop 0	thin tin 0 then 0 throw 0	about four 0 cow 0 above 0	car corn 0 arm 0 ran 0	caught laugh 0 bought 0 cow 0
snid	rug	sith	fout	pard	sauf

American students from a single school in Laredo, Texas in January, 1981. In both studies, the teachers were asked on a pre-arranged visit to supply several "average" Title I students. The teachers were told that the students would be administered a test to determine the effect on students' scores of various question-and answer formats from standardized tests. The students were administered the Format Familiarity Test in small groups of six except for the last section which was derived from the Woodcock and was administered individually. During test administration, each format section was preceded by abridged directions from the test from which it was derived.

Results

Study I

The cell and marginal means of percentage of correct answers are presented in Table 19. A two-way repeated measures ANOVA indicated a statistically significant difference in the percent of students scoring correctly on format types ($F = 32.41$, $p < .001$)(see Table 20).

Since the main effect for format was significant, the Newman-Keuls' Multiple Range Test for Differences Between Means was calculated. Group means are ranked from largest to smallest across the top of the table and from smallest to largest down the side. Pairwise differences between the means of all possible pairs of formats are presented in the matrix. The results are presented in Table 21. The differences between all pairwise comparisons of the means were statistically significant.

Study II

The cell and marginal means of percentage of correct answers for the Mexican-American students are presented in Table 22. A two-way repeated measures ANOVA indicated a statistically significant variation in the percent of students scoring correctly on format types ($F = 41.47$, $p < .001$)(Table 23).

Table 19.

Study 1: Cell and Marginal Means and Standard Deviations of Percentage Scores on the Format Test for Caucasian Students

	Format 1	Format 2	Format 3	Format 4A	Format 4B*
	91.90 (27.4)	55.41 (49.95)	46.40 (49.8)	64.87 (47.66)	50.90 (50.10)
Consonant Sounds sn	100.00 (00.0)	62.16 (49.77)	75.68 (43.50)	72.97 (45.02)	48.65 (50.71)
Short Vowel Sounds u	97.30 (16.43)	64.87 (48.40)	70.27 (46.34)	91.89 (27.07)	62.16 (49.17)
Consonant Digraph th	91.90 (27.67)	83.78 (37.37)	24.32 (43.50)	86.49 (35.07)	72.97 (43.92)
Vowel Digraph	91.90 (28.03)	40.54 (49.71)	18.92 (39.71)	62.16 (48.71)	51.35 (50.67)
Controlled Vowel ar	73.00 (45.02)	43.24 (49.17)	59.46 (49.44)	59.46 (49.17)	54.05 (50.54)
Variant Vowel au	97.30 (16.44)	37.84 (49.17)	29.73 (41.69)	16.22 (37.37)	16.22 (37.37)

*4A and 4B are derived from the Woodcock. 4A is scored correct if the target sound is pronounced correctly. 4B is scored correct only if the entire word is pronounced correctly as suggested by the manual.

Table 20

Source of Variance for Repeated Measures ANOVA with Caucasian Students

Source	Degrees of Freedom	Sum of Squares	F	Significance Level
Phonics	5	19.17	15.41	.001
Format	4	29.12	32.41	.001
Subjects	36	21.84		
P X F	20	20.48	7.37	.001
P X R	180	44.79		
F X R	144	32.34		
P X F X R	720	100.06		
TOTAL				

Table 21

The Newman-Keuls' Multiple Range Test of Differences Between Means of Format Types for Caucasian Students

		Format 1 .919	Format 4A .649	Format 2 .554	Format 4B .509	Format 3 .464
Format 3	.464	.455**	.185**	.090**	.045**	
Format 4B	.509	.410**	.140**	.045**		
Format 4A	.649	.270**				
Format 1	.919	0				

** = $p < .01$

Table 22

Cell and Marginal Means and Standard Deviations

Percentage scores for Mexican American Students

		Format 1	Format 2	Format 3	Format 4A	Format 4B*
		.57 (49.39)	.38 (48.52)	.41 (49.29)	.75 (43.80)	.58 (49.48)
Consonant Sounds sn	.57 (48.49)	.06 (24.97)	.84 (37.39)	.77 (42.50)	.81 (40.16)	.39 (49.51)
Short Vowel Sounds u	.66 (46.16)	.94 (24.97)	.42 (50.16)	.71 (46.14)	.71 (45.68)	.55 (50.59)
Consonant Digraph th	.55 (49.95)	.83 (32.39)	.52 (50.80)	.23 (42.50)	.74 (44.48)	.45 (50.59)
Vowel Digraph ou	.43 (49.48)	.87 (34.08)	.19 (40.16)	.00 (0.00)	.55 (50.59)	.55 (50.59)
Controlled Vowels ar	.53 (49.81)	.42 (50.16)	.13 (34.08)	.55 (50.59)	.84 (37.39)	.74 (44.48)
Variant Vowels au	.47 (50.14)	.38 (49.51)	.16 (36.89)	.19 (40.16)	.81 (40.16)	.81 (40.16)

*4A and 4B are derived from the Woodcock. 4A is scored correct if the target sound is pronounced correctly. 4B is scored correct only if the entire word is pronounced correctly as is suggested by the manual.

Table 23

Source of Variance for Repeated Measures ANOVA for Mexican American Students

Source	Degrees of Freedom	Sum of Squares	F	Significance Level
Phonics	5	5.16	11.77	.001
Format	4	16.48	31.44	.001
Subjects	30	82.41		
P X F	20	46.37	26.84	.001
P X R	150	13.14		
F X R	120	15.72		
P X F X R	600	51.83		
TOTAL		231.11		

Since the main effect for format was statistically significant, the Newman-Keuls' Multiple Range Test for Differences Between Means was calculated. The results are presented in Table 24. Significant differences between means at $p < .01$ are indicated with asterisks. The differences between the means of formats 1 and 5 and 3 and 2 are the only nonstatistically significant differences.

Discussion

The results of this study indicate that students score very differently on phonics items depending on the format used to present the items. This means that conclusions about how well a student has mastered phonics content will depend in part upon the format of the particular standardized test which is used.

Since raw score percentage correct is used as the dependent variable in these analyses, these differences would be meaningless if they were adjusted for in the test's norming procedures. In other words, a difference between a score of 50% correct on the format of Test A and 80% correct on the format of Test B would be unimportant if the 60 percentile for Test A was 50% correct, and the 60 percentile for Test B was 80% correct. As can be seen in Table 25, test norms from the tests used in the study do not account for the differences.

In support of previous research, the students performed best on those formats which did not require the use of skills and knowledge in addition to that required to answer the question. This is directly evidenced by the differences in scores obtained on formats 4A and 4B. Format 4A was scored correct if the target sound was pronounced correctly, whereas format 4B (the method recommended by the test publisher) was scored correct only if the

Table 24

The Newman-Keuls' Multiple Range Test of Differences Between Means for Format Types for Mexican American Students

		Format 4	Format 1	Format 5	Format 3	Format 2
		.742	.586	.581	.409	.376
Format 2	.376	.366**	.210**	.205**	.033	
Format 3	.409	.333**	.177**	.172**		
Format 5	.581	.161**	.005			
Format 1	.586	.156**				
Format 4	.742	0				

* = $p < .05$

** = $p < .01$

Table 25

Comparison of Normative Scaling of the Phonics Analysis Subtests of the CAT and SAT at the End of Second Grade

Percentile	SAT	CAT
	Percentage of raw score correct	
50th	73%	72%

entire word was pronounced correctly.. Format 4B required that the students know additional content besides that of the target content.

The Mexican-American students scored lowest on formats 2 and 3. These formats required an extra cognitive step in much the same manner as do recall items versus recognition items. Formats 1 and 4 require that the student simply pronounce or identify a word. Formats 2 and 3 require that the student discriminate between sounds within a word, isolate the sound, and identify that sound in another word. This is a far more complex task.

The Caucasian students also scored poorly on formats 2 and 3. They scored far better than the Mexican-American students on format 1 and worse on formats 4A and 4B. This suggests a disordinal interaction of format types with ethnicity. This may be evidence that a student's familiarity with the test's question format influences the student's scores on that test. All of the Mexican-American students used Spanish as their first language. Much of their educational experience may have been learning the correct pronunciation of English from written symbols. This is exactly the skill that format 4 samples and on which Mexican-American students did best. Correspondingly, much of the Caucasians' educational experience may have been recognizing the correct written symbols for words they already knew. This is the skill format 1 samples and on which the Caucasian students did best.

Since the main focus of this study was to investigate whether different format types yielded different results when testing identical content, the statistically significant main effect for phonics content and interaction with format type are not discussed further. The statistically significant interaction may be due, in part, to a sparse sampling of the domain of phonics items. This hypothesis could be investigated further if future research uses more questions from each phonics content category.

Although the results are based on a fairly small sample and the Format Familiarity Test made no effort to comprehensively sample the content domain, the results do indicate that what a student appears to know is confounded with the format of the test that is used. One explanation for this is that students may perform better on those tests which ask questions in the same way that is familiar to the student during instruction. In other words, using an unfamiliar format may well decrease the student's score by introducing an irrelevant variable and make it appear that the student has not mastered the material as well as he or she has.

School personnel have traditionally been concerned with selecting achievement tests that sample the content being taught in the curriculum. Not as much attention has been paid to selecting a test that asks questions about the content in the same manner as is done in the classroom. These results indicate that the format in which questions are asked is an important consideration for selecting an achievement test.

CHAPTER VI

EFFECTS ON STANDARDIZED ACHIEVEMENT TEST PERFORMANCE OF TRAINING TEACHERS, TRAINING STUDENTS, AND MOTIVATING STUDENTS

A second component which was added to the work scope of the "Refinements" contract was a study to investigate the effect on standardized achievement test performance of:

1) training teachers to follow standardized testing procedures;
2) training students in test taking skills; and
3) motivating students to perform well on the test.

During August and September of 1979, project staff visited a number of districts in Utah to observe standardized test administration. The purpose of these visits was to prepare for the project's data collection efforts the following spring, and to assist Salt Lake District in implementing Models A and B correctly. During these visits, staff members observed that many test administrators did not adhere to standardized test administration procedures and students appeared to be confused about what was expected of them and/or uninterested in the test. Conversations with teachers after the testing confirmed that many teachers were not adequately prepared to administer the tests and did not see the time spent on testing as worthwhile. As project staff were preparing the test booklets for automated scoring by the publisher, the number of stray marks, incorrectly followed directions (by both teachers and students) and generally sloppy or careless completion of the booklets made it clear that many people did not view the administration of standardized tests as an important activity.

These observations reinforced the concept that other variables in addition to what a student knows may be influential in determining a

student's test score. For Title I students the possibility that extraneous factors are confounding what a student appears to know is particularly important for two reasons:

1. Because standardized achievement test scores are assumed to reflect how much a student knows, the results of these tests are frequently used to make educational placement and programming decisions. If the test is not a valid indicator of what a student knows, placement and programming decisions based on the test may be incorrect.
2. All of the TIERS models utilize the results of standardized achievement tests to measure the impact of Title I. In this context "impact" is defined as what a student learns which he would not have learned had the Title I program not been offered. To be a valid measure of "impact", regardless of which model is used, the tests utilized must measure student's knowledge without being substantially confounded with other variables.

The importance of these issues to the objectives of the "State Refinements" project prompted the inclusion of a discussion of this added study in this Final Report. Although not included in the original work scope, the activities of this component were partially supported by project funds. Additional funds came from a small Student Initiated Research Grant ($6,801) from the Bureau of Education for the Handicapped (these funds paid for materials development, travel and current expense not included in the Refinements Project, and data collectors), and from contributed resources and time from Utah State University. The district in which this research was conducted was one of the districts included in the original "Refinements" project and many of the procedures and instrumentation developed for these activities were also utilized in

accomplishing the objectives of the Refinements contract. As a result of the symbiotic relationship between the original workscope and these extended activities, both components benefited, and the government received "more for their money" than had originally been planned.

The purpose of the study described in this chapter was to determine the effect of training students in test taking skills, training teachers in standardized test administration procedures, and reinforcing students for high performance during testing. The effect of treatment conditions on four dependent measures was examined using a true experimental design and data were analyzed by a three-way (2 X 2 X 2) analysis of variance. The dependent variables were student reading scores, student on-task behavior, teacher on-task behavior, and invalid test booklet marks.

Procedures

Subjects

Participants in the study (students and test administrators) were from the district's eight Title I schools (16 classes) and from four "similar" schools (eight classes). "Similar" schools were selected by the Title I director as those most closely approximating the Title I schools in mean IQ, previous achievement, achievement test scores, and income level. Analyses demonstrated that there were no statistically significant differences between the Title I schools and the similar schools on any of these variables (reported earlier in Chapter III - Table 2).

Permission to conduct the study in the Salt Lake City District was obtained from Dr. Darlene Ball, then Director of Title I for Salt Lake

City District, and Dr. Stanley Morgan, Research Coordinator for the district. The study was endorsed by Maurine McDonald, District Coordinator for Title I, and Dr. Cy Freston, Utah State Office of Education Specialist in Learning Disabilities. The research was approved by the Institutional Review Board at Utah State University.

Notice was sent to each parent (see Appendix 9 for a copy of the letter) describing the reinforcement procedures, the research rationale, and procedures for withdrawing their child from the study if they so desired. Personal individual contact was made with each principal in the 12 schools to explain the research procedures and secure the authority for treatment and data collection (see Appendix 9 for letters of support, approval and notification).

Students. Second grade classrooms were chosen to participate in the study because it is frequently at this level that group achievement tests are first encountered. A bad testing experience could negatively influence the students' attitudes toward future tests. Although all participating students attended district- selected Title I or "similar" schools, only 30% were actually Title I "target" students.

The selection criteria for classifying students as target were different, though related, for the Title I schools and the "similar" schools. In the designated Title I schools, the identification of a student as a Title I target student was based on students' spring performance on five key indicators:

1. Reading subtest (SAT).
2. Math subtest (SAT)

3. Language subtest (SAT)

4. Teacher checklist of behavior (locally developed)

5. Teacher evaluation of oral language skills (locally developed)

The students' scores on each indicator were assigned weighted point values. Students scoring below a locally determined criterion were identified as target students. In addition to meeting this selection criterion, students had to score below grade 1.8 on the reading or math subtests (SAT) that were given during spring of the previous year to be eligible for reading or math Title I programs.

Since the "similar" schools were not designated as Title I and there were no official target students in the classrooms, it was necessary to develop criteria for classifying a sample of students in the "similar" schools as unofficial "targets". The method used for identification was to select students with the lowest 19% of the scores on the Spring, 1980, SAT in the similar schools. The 19% cutoff was determined as the median percent of Title I students in the designated Title I classrooms. Hereafter, the term "target" will refer to the combined group of students from Title I and similar schools.

Of the 597 participating students, 323 (54%) were male; 180 (30%) "target" students and 46 (7.7%) were mainstreamed mildly handicapped special education students. All special education students were also identified as Title I students and were included in the analysis in the "target" group. The average number of students per classroom was 24, with a range of 18 to 33.

Test administrators. The standardized achievement test used as one of the dependent variables in the study was administered by 24

teachers--12 were Title I teacher leaders and 12 were regular classroom teachers. The teacher leaders were certified teachers who performed resource functions in the elementary schools (directing remedial instruction by the pull-out teachers and classroom teachers, testing students, programming instruction, and providing inservice training to district teachers).

During a group meeting, the study was explained to the teacher leaders who all agreed to participate. Teacher leaders were randomly assigned during the administration of the test for the study to one of the 12 regular second grade classrooms which had been selected earlier to have the test administered by a trained teacher.

Regular teachers were introduced to the study via a letter from the district Title I director and subsequently sent instructions regarding their particular duties during the testing. Untrained teachers were not told about the research variables, but were informed that they would have observers in their classrooms from time to time and that their students would be reinforced or trained. At the conclusion of the testing, the details of the study were provided to each of the teachers.

Test monitors. The Title I director selected 24 Title I pull-out teachers to serve as test monitors in all classrooms used in the study. The pull-out teachers were fully certified elementary teachers who worked a half-day under teacher leaders as instructors to remediate Title I students in reading and math. For the study, each monitor was randomly assigned to one classroom to assist the teacher (trained or untrained) responsible for the testing.

Treatment

The effect on student test scores and student and teacher behavior during testing of three separate factors was examined during the study: (a) reinforcing students for scoring higher on the spring test than would have been predicted from their fall test, (b) training students in techniques to increase their test taking skills, and (c) training teachers to administer the test using good testing practices and standardized testing procedures.

The design of the study was completely crossed; each treatment and control interfaced with other treatments and controls creating a 2 X 2 X 2 block, or eight cells. Each of 24 classes of students were randomly assigned to one of the eight cells of the experimental design shown in Table 26. Random assignment was made by drawing slips of paper containing concealed classroom identification numbers. The mean number of students per cell was 74.62 with a range from 67 to 85.

Reinforcement for students. Past procedures for motivating students to do well have included the use of many types of verbal and tangible rewards. Nickels were chosen as the reinforcement in this study for several reasons:

1. Students in Title I schools came from low income families in the community creating the strong possibility that money was highly valued.

2. Using money avoided the dietary problems of food (i.e., candy), special preferences of prizes, or the confounding personality variables of praise.

Table 26

Assignment of Classrooms to Treatment Groups

		Trained Teacher	Untrained Teacher	
Trained Students $n_{ts} = 291$	Reinforced	1[a] Bennion (25)[b] 12 Washington (23) 21 Backman (27) n = 75	7 Lowell (2[illegible] 15 Whittier (1[illegible] 24 Edison [illegible] n = 70	$n_{re} = 297$
	Unreinforced	5 Lincoln (29) 9 Parkview (18) 11 Parkview (24) n = 71	4 Jackson (27) 20 Hawthorne (22) 23 Edison (26) n = 75	
Untrained Students $n_{-ts} = 306$	Reinforced	8 Lowell (27) 19 Hawthorne (31) 22 Backman (27) n = 85	2 Bennion (23) 3 Franklin (21) 13 Washington (23) n = 67	$n_{-re} = 300$
	Unreinforced	14 Whittier (23) 17 Emerson (26) 18 Emerson (27) n = 76	6 Lincoln (33) 10 Parkview (22) 16 Whittier (23) n = 78	
		$n_{tt} = 307$	$n_{-tt} = 290$	

[a]The number preceding each classroom is a unique identification code.

[b]The number following each classroom indicates the number of students in each classroom.

3. Previous research has demonstrated that low IQ students perform better on individualized aptitude tests with monetary rewards than without (Rasmussen, 1973).

4. Money is easy to dispense and control.

5. Increments of nickels are equal and noticeable (i.e., the better the student does on the test, the bigger the pile of money received).

During Spring, 1980 district achievement testing, students in those classes which had been randomly assigned to the reinforcement condition were reinforced for doing better on group tests than was expected based on their Fall, 1979 test score. The amount paid to each student was based on the number of raw score points above an individually established base (expected) score. The procedures for determining the base score for each child are illustrated in Table 27 using hypothetical scores for three students.

Table 27

Examples of Procedures Used to Determine the Amount of Reinforcement Given to Students Based on Fall Test Scores

	Fall Testing Norms		Spring Testing Norms		Adjustment	Payment Criteria	Actual Spring Score	Reinforcement
	Raw Score	Percentile	Percentile	Raw Score				
Johnny	23	30	30	20	-5	15	25	$.50
Mabel	33	50	50	33	-5	28	27	-
Helen	40	70	70	40	-5	35	41	.30

For example, if a student had a raw score on the Fall test of 23 (or 33 or 40), this would be at the 30th (or 50th or 70th) percentile using Fall norms for that level of the test. If the student learned at the normal rate of students in the norming sample, the same percentile rank using Spring norms would be maintained on the Spring test. The 30th percentile (or 50th or 70th) using Spring norms for the level of the test used in the Spring is associated with a raw score of 20 (or 33 or 40). Each equipercentile score thus derived was adjusted by subtracting 5 points so that more students would be able to earn money.

In the example in Table 26, the payment criteria for Johnny (20 - 5 = 15) becomes the base score above which Johnny must achieve on the Spring test to be rewarded. If Johnny's actual Spring score was 25, then he would earn 10 nickels (25 - 15 = 10) or $.50. Using those procedures, payment criteria were individually established for each child. For students who didn't take the pretest (absent during Fall testing or transferred after Fall testing), teachers estimated a percentile rank for the pretest based on classroom performance.

Four subtests were selected as the units for reinforcement: Reading A, Reading B, Word Study, and Mathematics (Math Concepts plus Math Computations). Although students in the reinforced group were paid for performance on the Reading A, Reading B, and Word Study subtests, only data on Reading B and Word Study scores were used in the analyses. To convince students that they would really be given money if they did better than the payment criterion on the test, Reading A was administered to the students and reinforced on the day immediately preceding Reading B and Word Study. After observing the actual payment of money for their performance on Reading A, the students were more likely to believe that

they would get paid for doing better than the payment criterion on Reading B and Word Study.

Students in the unreinforced classrooms were paid for performance on Math subtests to prevent a negative reaction from the nonpayment group for not having a chance to earn nickels. All unreinforced classrooms received the reading tests before the math tests so that data used in making the reinforced vs. unreinforced comparison could be collected prior to delivering rewards for math performance.

The procedures for reinforcing students are outlined below.

1. Notification. Just before giving the directions to the subtest, the test administrator read this statement: "Today you will earn nickels for doing well on the next test. The higher you score, the more nickels you will get. Try very hard to do your best, and I (your teacher) will give you your money to take home this afternoon."

2. Scoring. On the day of reinforcement (same day as subtest), trained scorers traveled to each school with a test key, scored the subtest, computed the amount earned, and prepared an envelope of nickels for each rewarded student.

3. Payment. The classroom teacher presented the rewarded students with envelopes containing their earnings just before school ended for the day.

Training for test administrators and monitors. Training in appropriate test administration was provided by the investigator to the 12 Title I teacher leaders (test administrators) and to 12 randomly selected Title I pull-out teachers (test monitors) prior to the group achievement testing in the Spring. The training program for each group of teachers is outlined below.

Teacher leaders were selected to receive training instead of regular classroom teachers for the following reasons:

a) they fully supported the study and had expressed a desire to bring a test preparation program into the district;

b) because they had fewer habitual reactions to particular students than would the regular classroom teacher, it was thought they would be more consistent and dependable in carrying through with appropriate behaviors learned during the training; and

c) teacher leaders were chosen over nonteaching personnel because they had some familiarity with the students and with proper group testing procedures.

To prepare for the test administration training, a videotape was constructed that depicted actual classroom scenarios of correct and incorrect methods of giving tests. Second grade students from Wellsville, Utah, and their teacher agreed to act in contrived situations to portray the results of appropriate and inappropriate test administration. Permission for filming was obtained from the teacher, principal, and parents, and all had an opportunity to view the final product (see Appendix 10 for approval forms).

Displayed on the film are both the right and wrong methods for giving tests. Previous observations and research were used to compile the script and included:

1. Preparing students for the test.
2. Arranging the testing room.
3. Entering the testing room.
4. Distributing test materials.

5. Giving directions.
6. Monitoring the students.
7. Using an aide.
8. Finishing the test.
9. Providing assistance to students.
10. Dealing with unexpected events.
11. Pacing.
12. Obtaining group response.

Training took place in the Salt Lake City District Offices two weeks before the district testing. Two sessions, on the afternoons of April 15 and 17, were held from 1:30 to 3:30. Training content covered general test administration, on-task teacher behavior, the Stanford Achievement Test (SAT), and the use of test monitors. Procedures for obtaining group response and for explaining test format were included in the training. Special attention was given to unexpected testing problems and teachers were instructed on appropriate responses. For example, if a child needs to use the bathroom during a timed test, the administrator should record the time absent and extend the test by that amount for the individual student.

The Quality of Test Administration Checklist (described in detail in Chapter IV) was explained prior to seeing the videotape and was used as an observation guide during the film. Differences in student behavior under incorrect and correct test administration were then discussed. Directions for the SAT Practice Test, Level II, were discussed and each trained teacher was asked to administer the practice test to their assigned classroom. Copies of the practice test were provided to the teachers, and they were instructed to use the practice test as a teaching device and not as a test.

The Title I director selected 24 Title I pull-out teachers for use as test monitors. They were randomly assigned to classrooms and all those who were assigned to classes which had trained test administrators were given training. On April 7, 1980, letters were sent to the selected pull-out teachers informing them of their participation in the upcoming Spring testing and indicating the training schedule (for the 12 who were to be trained).

Training provided to pull-out teachers in test monitoring resembled that given to the teacher leaders but was condensed to 2 1/2 hours on Wednesday, April 16, 1980. Specific classroom testing problems and methods for alleviating them were discussed. For example, a student who is constantly off task could be given physical prompts (i.e., a hand on the shoulder, assistance in moving finger from item to item during timed tests). Like the test administrators, monitors were encouraged to actually take the test as if they were students prior to the time the test was administered to the students.

Training for students. Twelve classrooms were randomly selected to participate in student training in test-taking skills consisting of coaching in answering specific types of items, training in testwiseness (i.e., how to guess or deduce answers), and answering sample items on the SAT Practice Test, Level I. Students were trained by the investigator in their own classrooms for one hour during the morning of a regular school day 1 to 2 weeks before the actual testing. The classroom teacher was not present during the training and each session followed the same format. The training schedule, a complete copy of training procedures, and the Practice Test are available on request.

The activities and topics covered during the student training included:

1. Purpose of the test.
2. Group response.
3. Positive atmosphere.
4. Rules for correct testing behavior.
5. Machine scorable answer forms.
6. Unusual directions.
7. Practice test.
8. Multiple choice items.
9. Rest period.

At the end of 50 minutes, the lowest five performers identified during the training were given 10 additional minutes of individual help, while the others completed some puzzles.

Dependent Measures

Four dependent variables were measured to assess the effectiveness of the various intervention conditions: reading test scores, student on-task behavior, teacher on-task behavior, and test booklet marks. Instruments and procedures for collecting data and programs for training observers to record the data were developed during the study. The following sections describe the dependent variables and the instruments used to collect data.

Test scores. The Stanford Achievement Test (SAT), Primary Level II, Form A (Madden, et al., 1972), used for the Spring test in grade two to measure the impact of the interventions on students' reading scores.

Level II is both a norm-referenced and objective-referenced test designed for group administration to assess skill development in vocabulary, reading, mathematics, spelling, word study, and listening comprehension. The manual for the SAT contains the general directions for administering the test as well as the specified verbal instruction to be read to the students for each subtest. Reliability data for the SAT consists of split-half estimates and KR-20 coefficients. Reliabilities for all subtests at all levels on all forms range from .65 to .97 with the majority between .85 and .95. Evidence for the validity of the tests consists of two types of information: (a) an increasing difficulty of items with higher grade levels, and (b) a moderate to high relationship with previous SATs and with the current and previous Metropolitan Achievement Tests.

Based on test scores from September, 1979, Level II was selected to use in testing second grade during Spring, 1980, to avoid floor and/or ceiling effects. The entire test (excluding optional subtests) was administered in four days to students in each treatment group according to directions specified in the manual. Make-up tests were given to absent students from May 2 through May 5. During the scheduled testing, subtests were given each morning, the first section administered after morning exercise followed by a recess and then the second session.

Only one day of the four-day test was utilized for data collection. A combined score of the Reading B (RB) and Word Study (WS) was selected as the most suitable dependent measure because of the variety in item format, item content (comprehension and phonics), and subtest sequence in the testing schedule. Both timed tests, where students move at their own pace, and teacher-directed tests, where the class moves as a unit, were

part of RB and WS. Reliability coefficients established during the norming were .96 for RB and .94 for WS.

The order of administering RB and WS was varied among the classes, but both subtests were always given on the same day. Since Vocabulary and Reading A subtests precede RB and WS, no data were collected on the first day of testing, and all students had some testing experience before taking RB and WS.

Test booklets were prepared for machine scoring by erasing all irrelevant marks and darkening light circles. Tests were scored by the Utah State Office of Education. A combined raw score, RBWS (113 total) raw score points was used as the test score dependent measure.

Student on-task behavior. A review of the literature and previous observations contributed to a list of appropriate behaviors most conducive to producing high levels of attention to academic tasks. The definition of student on-task behavior used in this study was described earlier in Chapter IV and definitions, examples, and nonexamples of acceptable activity under both teacher-directed and student-directed (timed tests) test taking are provided in Appendix 5. To illustrate the application of the definition, suppose a girl were twisting a shirt button with her fingers. This behavior (see "body movement, playing with clothes") is on-task if she is not looking at the button (see "looking, at test paper") but off-task if she is looking (see "looking, at clothes").

Student on-task behavior and teacher on-task behavior (see below) were both recorded on an interval form developed, field tested, and revised for the study as explained in Chapter IV (see Appendix 5).

Prior to actual data collection, an interrater reliability coefficient was obtained to establish instrument reliability over six trials completed after the last revision. Reliability was computed for each subject separately by the equation:

$$\text{Interrater Reliability} = \frac{\text{Number of agreements}}{\text{Number of agreements} + \text{Number of disagreements}} \quad (1)$$

where: Number of agreements = intervals that two observers record the same mark, and Number of disagreements = intervals that two observers record different marks.

For the field test, three graduate students were paired for six different trial observations and collected data on five students and one teacher during each observation. Table 28 contains the reliability coefficients for students and teachers for each field test trial. Mean coefficients of .878 for student on-task behavior and .854 for teachers were obtained by averaging coefficients across trials.

TABLE 28

Interrater Reliability Coefficients for Trial Observations of On-Task Behavior

Subjects Observed	Trials					
	1	2	3	4	5	6
Students (mean)	.865	.875	.712	.846	.985	.984
Teacher	.625	.844	.911	.867	.941	.933

During the actual research study recordings of student on-task behavior were made by trained observers during Reading A (RA), the RB and WS subtests. Teachers were asked to identify the five Title I or lowest achieving students in their classroom. Observations were made on the same students for both subtests. On the observation form, observers identified the students by a physical characteristic and used a tape recorded beeper to move from block to block to record "on task" or "off task". Recording began when the test administrator started reading the directions and ended when the subtest was completed. Five second intervals consisted of 3 seconds to observe and 2 seconds to record. Observers watched each child for 4 consecutive intervals or 20 seconds (4 intervals x 5 seconds = 20 seconds) before moving to the next student.

Data was recorded on five students and one teacher (six subjects) for 2 minutes (6 X 20 seconds) before repeating the cycle. The average observation time for the RA, RB, and WS subtests was 29.7 minutes. Since testing time varied across classrooms, the numerical unit chosen for analysis was the mean percent of on-task behavior per RB and WS observation. Percentages for each student were computed by the equation

$$\text{Percent of On-task Behavior} = \frac{\text{Intervals On-task}}{\text{Total Intervals Recorded}} \qquad (2)$$

The mean percent was the average percent across the five students observed and the two observers if the observation was paired.

Teacher on-task behavior. Standardized testing procedures listed in the SAT teacher's manual and preliminary observations formed the basis for defining teacher on-task behavior (explained in Appendix 5). Actions consistent with attending to student behavior at all times (while directing the test administration under standardized conditions) are included as examples of on-task behaviors. For example, a teacher is on task when reading directions but off task when talking to the entire class during a timed test (TT).

The interval recording form described above for recording student on-task behavior was also used by trained observers to collect data on the teacher. The last set of four 5-second intervals on the recording form was employed to watch the teacher (20 seconds every 2 minutes). The numerical unit used in the analysis was the percent of on-task behavior (see Equation 2).

Test booklets. Marks that would invalidate answers when scored by machine were identified as an indication of inappropriate student testing behavior. Invalid marks occurred when students drew pictures, skipped items, filled in more than one answer, tore booklets, erased too hard, used crayons or ink, or wrote too lightly. Examples and procedures for data collection are described more fully below. The numerical unit used in the analyses was the percent of items or test booklets with errors.

Observing and Reinforcing

Personnel hired to observe and deliver reinforcement were mothers of Title I children contacted through the Title I Parent Advisory Council. During an orientation meeting held on April 21, 1980,

observation procedures were explained and 17 people contracted to work. The pay rate of $4.00 per hour included travel from home and data collection. Three types of work were offered to the observers: collecting data on testing behavior, scoring and reinforcing reading and math subtests, and collecting data on test booklets. The procedures for training observers and implementing each activity are described below.

Collecting data on test behavior. Seventeen observers attended all the training and collected data on student and teacher testing behaviors. Two major training strategies were involved: (a) training prior to going into the classroom during which observers familiarized themselves with the system and practiced using the system by scoring a videotape of unrehearsed testing scenes, and (b) practice in the classrooms. Two training sessions were held on April 21 and 25 to explain the subtests to be given during the observation time, the student and teacher on-task behavior definitions, the recording form, the observation procedures, and the observation schedule.

Observers were trained to collect data by intervals moving from cell to cell on the form at a signal from a tape recording that indicated when to observe (3 seconds) and when to record (2 seconds). Portable tape players were equipped with earphones for two people to use simultaneously, facilitating interrater reliability calculation.

Recording started when the teacher began the directions and observers marked each cell for on-task (1) or off-task (-) behavior. Each of the five students and the one teacher was observed for 20 consecutive

seconds, or four cells, during each 2-minute block of time. Observers computed percent on-task by dividing the number of "1" marks by the total number of intervals. All computations were checked by a second person and errors adjusted.

Following each videotape practice, data were checked against the standard (correct recording), discrepancies discussed, and forms collected for computing reliability. Data collection from videotaped scenes was practiced on both training days and during a short meeting after the classroom practice. Reliability was computed from data comparing each observer with the standard using Equation 1. The obtained coefficients for each observer ranged from .86 to .93 with an average of .90.

Observers were assigned in pairs to classrooms to practice data collection on April 28. During the first subtest (vocabulary), observers watched the testing to get a "feel" for the classroom situation and recorded no data. Observers did record behavior on the interval form during the second subtest, Reading A, as practice and to compute reliability. Data collected during the classroom practice obtained an interrater reliability of .88 using paired observations with Equation 1. In addition to using Equation 1, pairwise correlations of the paired observations for the five student on-task percentages were computed across observers. Mean correlations were RA = .84 (SD = .20), RB = .89 (SD = .12), and WS = .81 (SD = .15).

Actual observations began on Tuesday, April 29, the second test day and continued through Thursday, May 1. Observers were randomly assigned to classrooms administering RA, RB, or WS and collected data alone (31 observations) or in pairs (38 observations), depending on the time and

day. One observer functioned as a substitute to replace absentees. Data from all the paired observations were used to compute interrater reliability using Equation 1. Reliability ranged from .74 to .97, with a mean of .86 for WS, .91 for RB, and .88 for RB and WS combined.

Test scoring and reinforcing. Twelve observers were trained for scoring subtests and computing money earned as reinforcement for performance above payment criterion. Students in the reinforced group were paid money for scores on Reading A, Reading B, and Word Study, and students in the unreinforced group (control) were paid for Math scores. Scorers were trained at the Salt Lake City District Office from 2:00 to 4:00 p.m. on Monday, April 28, 1980. Training consisted of practice in scoring samples of all subtests and computing money earned based on the results. Scores for correct and incorrect answers were recorded on a sheet separate from the booklet and totaled. This total (the actual score) was placed on a Reinforcement Record that contained precomputed individual cutoff scores above which the students had to score to earn money. For students whose names were not on the Reinforcement Record, the raw score corresponding to the 50th percentile was inserted as the cutoff score. The difference between the actual score and the cutoff score for each child was the basis for payment. All scorers obtained 100% correct on each sample test score and computation.

Scorers were randomly assigned to classrooms scheduled for reinforcement and reported to the school during lunch the same day that the appropriate subtests were administered with rolls of nickels, envelopes, an answer key, and a Reinforcement Record. Test booklets were scored at the school, reinforcement computed, and the appropriate number of nickels

was placed in an envelope for each rewarded student and left with the teacher. Tests taken from April 29 through May 1 were scored by the scorers. Make-ups were handled by the teachers, and money was delivered to the schools based on their reported needs.

The mean amount of money earned per student under reinforcement conditions (RBWS) was 91¢ for target students and 45¢ for non-target students. Under unreinforced conditions, target students were paid a mean of $1.13 per target student and 80¢ per non-target student scores. Of the 597 students, 519 (87%) received money during the testing week.

Test booklet data collection. Seven observers collected data from test booklet marks and prepared them for machine scoring. Interrater reliability data was collected at random intervals when pairs of observers were periodically assigned to do the same booklet and data were compared. Correlations were computed by observer pairs on the numbers of errors recorded per student and were used as reliability coefficients. An average interrater reliability coefficient of .99 was obtained from a range of coefficients of .93 to 1.00.

During the training, observers practiced scoring sample tests containing all types of violations. A list of errors, training samples, and the data collection form are provided in Appendix M. One recording form was used for each student, and data from each page were recorded on separate lines. Information collected on booklet covers was the number of covers that had blank circles, or had wrong circles darkened. Other data were number of booklets that had been erased (prepared for machine scoring) by the teacher, number of items that required erasing or

darkening, and number of items with no answer, more than one circle, or the wrong answer format. As part of their duties, observers were required to prepare the booklets for machine scoring after recording data. To prepare the pages, extraneous marks were erased, light circles darkened, cover information corrected, and answers recorded properly.

Results and Discussion

To assess differences attributable to the various factors included in the experimental design, a Univariate Analysis of Variance (ANOVA) was performed for each of the dependent variables.[3] As described in the Procedure section, the independent variables consisted of training students (TS), training teachers (TT) and reinforcing students (RE). Dependent variables were reading test scores, student on-task behavior during testing, teacher on-task behavior during testing, and students' invalid marks on test booklets.

In the analyses, the mean classroom score on each variable was used as the unit of analysis because entire classes were randomly assigned to the experimental conditions and training was applied to the class as a whole. The following sections describe the results from the analysis for each dependent variable.

---

[3] To protect against inflating the Type I error rate, Multivariate Analysis of Variance (MANOVA) is sometimes suggested when multiple dependent measures are being examined. However, since only four dependent variables were examined in the study and since univariate analyses are typically the second step in a MANOVA, it was concluded that a MANOVA would make the analysis unnecessarily complex without contributing any significant advantages.

Test Scores

A three-way ANOVA was used to determine if students' composite reading subtest scores (RBWS) were statistically significantly different under various treatment conditions. Additionally, the standardized mean differences between treatment and nontreatment were examined to determine the educational significance of the findings.

All students. Results for the combined RBWS score for all students are presented in Table 29. Statistically significant main effects were found on TT, F (1,16) = 3.48, $p < .10$ and RE, F (1, 16) = 16.77, $p < .001$. Means for RBWS scores (presented in Table 30) indicate that students receiving reinforcement ($\bar{X}$ = 87.53) obtained higher RBWS scores than students who were not reinforced ($\bar{X}$ = 77.54) and students with trained test administrators had higher RBWS scores ($\bar{X}$ = 84.82) than students with untrained administrators ($\bar{X}$ = 80.26).

Even more important than the statistical significance of these differences, however, is the educational significance. A number of approaches have been suggested for estimating the educational or practical significance of an observed difference. The Joint Dissemination Review Panel (JDRP) suggests an approximate rule of thumb for most education measures. If the difference between two groups is larger than 1/3 of a standard deviation, the difference can be considered to be educationally significant[4] (JDRP, 1977). Others have suggested

[4]This way of depicting differences between groups has been referred to as an "effect size" (ES) by Glass (1977). Computationally, it is derived by the following equation: $\frac{\bar{X}_T - \bar{X}_C}{SD_C}$. In the remainder of this section, effect sizes will always be computed with the standard deviation of scores using individuals as the unit of analysis.

Table 29

Summary of Three-Way ANOVA of Treatment Conditions on RBWS Scores for All Students

Source	SS	df	SM	F
Trained Students	62.05	1	62.05	1.74
Trained Teachers	124.17	1	124.17	3.48*
Reinforced Students	598.90	1	598.90	16.77**
TS X TT	154.79	1	154.79	4.34*
TS X RE	28.76	1	28.76	.81
TT X RE	37.93	1	37.93	1.06
TS X TT X RE	10.92	1	10.92	.31
Error	591.29	16	35.71	
Total	1588.80	23		

*p < .10
**p < .001

measures such as ω^2 (omega squared) as an indication of how much variance in the dependent measure is accounted for by a particular independent variable (Hayes, 1973). The computation of ω^2

$$\omega^2 = \frac{SS_{between} - (J-1)MS_{error}}{MS_{error} + SS_{total}}$$

is an effort to further quantify the strength of the relationship between an independent and a dependent variable. Although Glass and Hakstain (1969) showed how the interpretation of ω^2 can be problematic in certain situations because it may underestimate the importance of the relationship, ω^2 does provide additional information which can be used in determining the importance of differences detected with Analysis of Variance techniques.

As noted in Table 30, the standard deviation for RBWS when students are the unit of analysis is 21.49. Thus, the effect size of RE is .46 ([87.53 - 77.54] / 21.49 = .46) or almost one half a standard deviation unit. Estimating the strength of the relationship using ω^2 indicates that whether or not students were reinforced, accounts for 34.7% of the variance in their test scores. The effect size for the trained teacher (TT) condition is .21 or approximately 1/4 of a standard deviation, and computation of ω^2 indicates that this factor is accounting for approximately 5.4% of the variance. The TS condition has an effect size of -.15 and accounts for only 1.6% ([62.05 - 35.71] / [35.71 + 1588.8]) of the variance in students' test scores.

Further interpretation of the magnitude of the .46 (RE) and .21 (TT) effect sizes results from examining the range and skewness (-.62) of the frequency distribution of the test scores (RBWS). An 85 point range of

Table 30

Mean RBWS Scores and Standard Deviation for All Students in Each Treatment Condition

		RE	-RE	
$\bar{X}_{TS} = 80.93$ $SD_{TS} = 9.10$	TS — TT	(01) 97.87 (10.78) (12) 83.90 (25.75) (21) 84.29 (18.08) $\bar{X}$ =88.69 SD =7.96	(05) 70.41 (20.12) (09) 76.39 (22.34) (11) 71.13 (22.21) $\bar{X}$ =72.64 SD = 3.26	$\bar{X}_{TT} = 84.82$ $SD_{TT} = 9.10$
	TS — -TT	(07) 96.56 (13.29) (15) 76.94 (14.82) (24) 82.56 (25.08) $\bar{X}$ = 85.35 SD =10.10	(04) 70.26 (26.16) (20) 80.31 (22.88) (23) 80.54 (19.01) $\bar{X}$ = 74.04 SD = 5.87	
$\bar{X}_{-TS} = 84.14$ $SD_{-TS} = 7.48$	-TS — TT	(08) 96.27 (10.92) (19) 93.79 (15.35) (22) 90.27 (17.45) $\bar{X}$ = 93.44 SD =3.02	(14) 80.52 (20.78) (17) 88.58 (17.97) (18) 84.33 (22.70) $\bar{X}$ =84.48 SD = 4.08	$\bar{X}_{-TT} = 80.26$ $SD_{-TT} = 7.09$
	-TS — -TT	(02) 83.71 (21.27) (03) 76.47 (22.87) (13) 87.77 (23.31) $\bar{X}$ = 82.65 SD =5.72	(06) 73.36 (22.11) (10) 73.91 (22.27) (16) 80.77 (16.90) $\bar{X}$ = 76.01 SD = 4.13	
Using classrooms as the Unit of Analysis: $\bar{X}$ = 82.54 SD = 8.31		$\bar{X}_{RE} = 87.53$ $SD_{RE} = 7.44$	$\bar{X}_{-RE} = 77.54$ $SD_{-RE} = 5.88$	Using students instead of classes as the Unit of Analysis: $\bar{X}$ = 82.59 SD = 21.49

NOTE: Numbers before each cell entry identify the classroom. Numbers in parentheses after each entry are the standard deviations for each classroom completed by using the individual student as the Unit of Analysis.

scores from a 113 maximum indicates a large variance and a high probability of a small effect size when comparing classroom means with the individual student standard deviation. The negative skewness occurs because the median test score for all students was 88.15 with the bottom 50% of the students scoring from 28 to 88 (61) points and the top 50% scoring from 89 to 112 (24 points). The fact that half of the students received scores in only 28% of the range and 20% of the students scored within 10 points (from 103 to 113) of the maximum possible (113), suggests a ceiling effect on the test results. Consequently, although substantial increases are attributable for these two treatment conditions, the full impact of the reinforcement and trained teacher conditions may not have been demonstrated because the top students were obtaining scores near the maximum.

The mean raw score for RB and WS, reported in Table 31, show a difference between RE treatment and nontreatment groups of 4.5 for RB and 5.5 for WS. Translating the raw score to percentiles (Table 31) a difference under RE in percentile rank is 8 points (54 - 46) for RB and 14 points (62 - 48) for WS.. The difference between RE treatment and nontreatment groups in grade equivalence is over a half a year for WS (3.6 - 2.8 = .8).

Taken together, the information yielded by calculating the effect size and ω^2 for each of the main effects indicates that whether or not students are reinforced is educationally as well as statistically significant in accounting for the results of group administered standardized achievement tests. Motivating students to try hard on the test seems to be particularly important. Training test administators appears to be a moderately important variable, while training students (at least as it was done in this study) does not appear to be important.

Table 31

Raw Score, Percentile, and Grade Equivalent for Mean RB and WS Scores for Treatment and Non-Treatment Groups Under RE and TT Conditions

	RB		WS	
RE	Treatment	Non-treatment	Treatment	Non-Treatment
Raw Score	35.49	30.99	52.04	46.56
Percentile	54	46	62	48
Grade Equivalent	2.9	2.7	3.6	2.8
TT				
Raw Score	34.59	31.88	50.21	48.28
Percentile	54	48	56	52
Grade Equivalent	2.8	2.7	3.3	3.7

Given the way that students were randomly assigned to treatment conditions, these results suggest that the motivational level of the student and the conditions of test administration are causally related to the score obtained by the student.

Although not statistically significant, untrained students did receive higher RBWS scores ($\bar{X}$ = 84.14) than trained students ($\bar{X}$ = 80.93) (Table 30). This can be interpreted to mean that differences this large or larger would be obtained more than 10 times out of 100 if two samples of this size were randomly drawn from the same population. This evidence suggests that training students with the type of test-taking package described in this study will not influence test scores of second grade students from Title I classrooms in a metropolitan area.

A statistically significant two-way interaction (TT by TS) was found, F (1,16) = 4.34, $p < .10$ (Table 29). These results indicate that training students in test-wiseness is influenced by the status of the test administrator (whether trained or not). Similarly, the results of training test administrators is influenced by the degree of test taking training provided to the students.

Graphing the interaction (see Figure 1), it can be seen that trained students had higher RBWS scores when the test was administered by untrained teachers ($\bar{X}$ = 81.2) than when the test was administered by trained teachers ($\bar{X}$ = 80.7). Conversely, untrained students scored higher under trained teachers ($\bar{X}$ = 89.0) than under untrained teachers ($\bar{X}$ = 79.3). These results indicate that the training provided to students was not an effective agent in increasing scores and may even have had a slight detrimental effect when coupled with the effect of having a trained teacher administer the test. That is, trained teachers

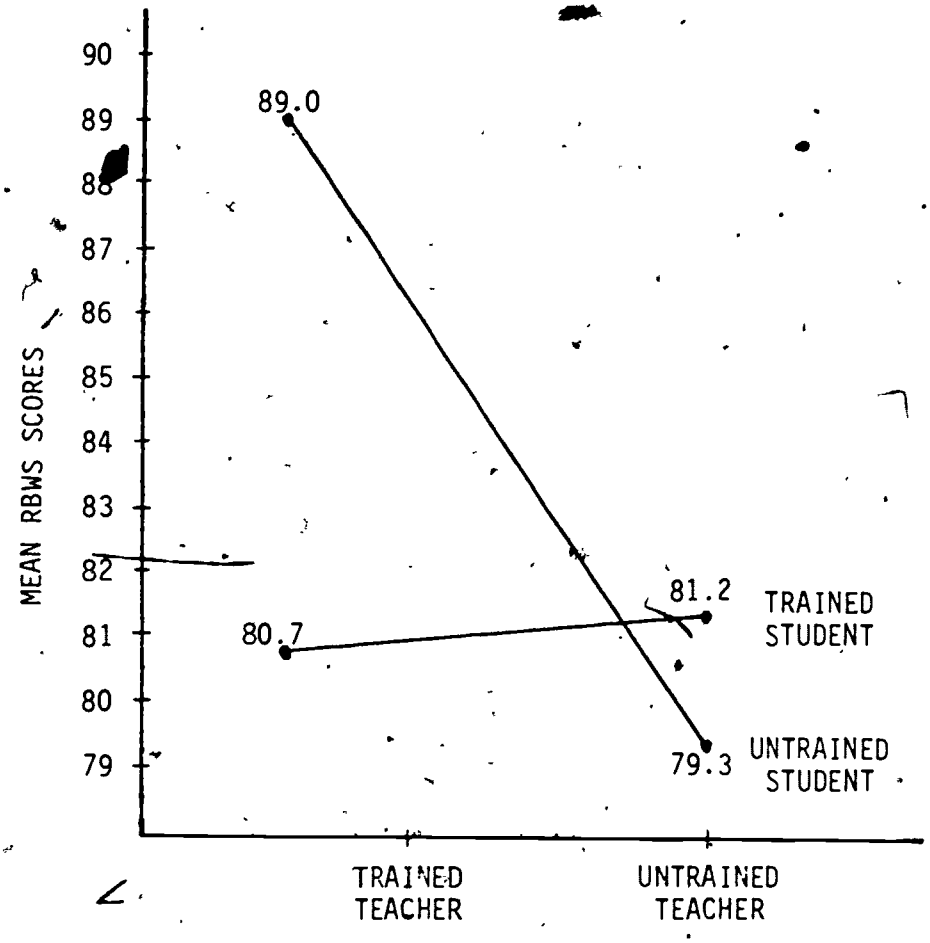


Figure 1. The graph illustrates the interaction of TT X TS for all students. Numbers in the graph are the mean RBWS scores obtained by classrooms under specified conditions.

were more influential in raising scores of the untrained students than of the trained students.

One possible explanation for the interaction is that trained students were too confident about testing skills and did not attend well to the actual testing task. Another possible explanation for the lower scores of trained students is that training was not effective and even confused the students because the trainer and the trained test administrator were strangers, and the untrained test administrator was their regular classroom teacher. These findings suggest that higher test scores may be obtained if the test administrator is familiar to the students, is trained in appropriate testing procedures, and provides students with reinforcement (motivation) for trying to do well.

Pretest/no pretest students. Analysis of covariance (covarying on a pretest taken during Fall, 1979) was originally planned for this study to improve the statistical power of the design. Scores on the pretest were available for 428 of the 597 students in the study, consequently, the use of ANCOVA would have resulted in only 72% (428/597) of the students being included in the analysis. To support the use of ANCOVA, students with the pretest scores should have been representative of students without pretest scores. A breakdown of test scores (Table 32) was prepared to assess whether posttest scores for students with pretest scores were any different from the posttest scores of students for whom pretest scores were not available.

The posttest means and standard deviations for the two groups are reported for RB, WS, and RBWS for each treatment condition. In every condition, the group for whom pretest scores were available had higher

Table 32

Mean Test Scores and Standard Deviations for Students With Pretest Scores and Students With No Pretest Scores

Treatment Groups	Number		RB		WS		RBWS	
	Pretest	No Pretest	Pretest	No Pretest	Pretest	No Pretest	Pretest	No Pretest
Reinforced Students	207	70	36.10 (10.5)	34.81 (10.2)	52.43 (10.35)	52.36 (10.6)	88.53 (19.6)	87.17 (19.6)
Unreinforced Students	221	71	32.34 (10.7)	26.24 (13.3)	47.50 (10.8)	43.27 (12.3)	79.84 (20.3)	69.51 (24.5)
Trained Teachers	223	72	35.32 (10.3)	32.79 (11.3)	50.70 (11.0)	49.03 (11.7)	86.03 (20.2)	81.82 (21.9)
Untrained Teachers	205	69	32.90 (11.2)	28.10 (13.5)	48.99 (10.7)	46.48 (12.9)	81.89 (20.5)	74.82 (25.3)
Trained Students	204	68	33.06 (10.8)	29.51 (13.3)	49.55 (11.0)	45.99 (13.0)	82.61 (20.7)	75.50 (25.1)
Untrained Students	224	73	35.17 (10.7)	31.41 (11.9)	50.19 (10.7)	49.45 (11.5)	85.35 (20.1)	80.86 (22.5)
All Groups	428	141	34.16 (10.7)	30.50 (12.6)	49.89 (10.8)	47.78 (12.3)	84.04 (20.4)	78.28 (23.8)

mean posttest scores than the group for whom pretest scores were not available. The mean RBWS (posttest) score for all treatment conditions was 84.04 for the pretest group and 78.28 for the no pretest group. The posttest score difference between the pretest group and the no pretest group was statistically significant, t (595) = 2.77, $p < .01$.

Due to the posttest score differential between the pretest and the no pretest group, it was concluded that students for whom pretest scores were available were not representative of the entire sample. Therefore, using ANCOVA would have biased the results and was deemed inappropriate for the study.

Teacher Behavior

Observational data of teachers' on-task behavior during testing were collected during the Reading B and Word Study subtests. A three-way analysis of variance was used to analyze the data across treatment conditions. Results presented in Table 33 show statistically significant main effects for trained teachers, F (1,16) = 36.34, $p < .001$.

As indicated in Table 34, the mean percent of on-task behavior for trained teachers ($\bar{X}$ = 72.9) was statistically significantly higher than for untrained teachers ($\bar{X}$ = 25.5). Although not statistically significant, means for the other treatment groups ($\bar{X}_{TS}$ = 50.7; $\bar{X}_{RE}$ = 54.9) were higher than the non-treatment group ($\bar{X}_{-TS}$ = 47.7; $\bar{X}_{-RE}$ = 43.5). Individual teacher percentages of on-task behavior ranged from 0% to 100%.

Table 33

Summary of Three-Way ANOVA of Treatment Conditions on Observations of Teacher Behavior During RBWS

Source	SS	df	MS	F
Trained Students	31.28	1	31.28	.09
Trained Teachers	13357.60	1	13357.60	36.34*
Reinforced Students	704.17	1	704.17	1.9
TS X TT	1.08	1	1.08	.00
TS X RE	13.95	1	13.95	.04
TT X RE	85.50	1	85.50	.23
TS X TT X RE	2.41	1	2.41	.01
Error	5880.49	16	367.53	
Total	20076.49	23		

*$p < .001$

Table 34

Mean Percent of Teacher On-task Behavior During RB and WS
By Treatment Condition

		Trained Teachers	Untrained Teachers	
Trained Students $\bar{X}_{TS}$ = 50.7 SD_{RS} = 29.4	Reinforced Students	98.8 68.1 70.1 $\bar{X}$ = 79.0 SD = 17.2	23.7 60.1 24.2 $\bar{X}$ = 36.0 SD = 20.9	$\bar{X}_{RE}$ = 54.9 SD_{RE} = 29.0
	Unreinforced Students	82.9 67.3 58.0 $\bar{X}$ = 69.4 SD = 12.6	35.4 8.4 11.0 $\bar{X}$ = 18.3 SD = 14.9	
Untrained Students $\bar{X}_{TS}$ = 47.7 SD_{TS} = 31.4	Reinforced Students	85.8 52.1 85.9 $\bar{X}$ = 74.6 SD = 19.5	12.4 60.6 17.4 $\bar{X}$ = 30.1 SD = 25.5	$\bar{X}_{RE}$ = 43.5 SD_{RE} = 30.0
	Unreinforced Students	81.3 70.6 53.5 $\bar{X}$ = 58.2 SD = 31.6	2.0 5.0 46.0 $\bar{X}$ = 17.7 SD = 24.6	
Using individual observations as the Unit of Analysis $\bar{X}$ = 51.7 SD = 29.66		$\bar{X}_{TT}$ = 72.9 SD_{TT} = 14.4	$\bar{X}_{TT}$ = 25.5 SD_{TT} = 20.6	

Using individual observations as the unit of analysis, the overall standard deviation of teacher on-task behavior is 29.66. The effect size (Equation 3) of TT is 1,6. The strength of the relationship between trained and untrained teachers (estimated by computing ω^2 - Equation 4) indicates that 63.5% of the variance in teacher on-task behavior is accounted for by whether or not teachers were trained.

Educational significance is a bit more difficult to establish here since the data collection instrument is not a normed test with which we have broad experience. However, these findings indicate that teachers who were trained in appropriate test administration techniques were demonstrating those skills substantially more frequently than untrained teachers.

Student Behavior

Data for student on-task behavior were analyzed by ANOVA across treatments and the findings presented in Table 35 show no satistically significant main effects or interactions. Differences this large would be obtained between two groups more than 10 times in 100 if samples of this size were randomly drawn from the same population.

Based on these findings, it appears that treatment conditions (RE, TS, and TT) did not influence the degree of on-task behavior displayed by the lowest five achievers in each classroom. Means presented in Table 36 show very little difference between treatment and non-treatment groups for each factor. In order, treatment and non-treatment on-task means were 72.9 and 73.9 (TS), 75.6 and 71.2 (RE), and 75.5 and 71.2 (TT). Individual student percentages of on-task behavior ranged from 35% to 100%.

Table 35

Summary of Three-Way ANOVA of Treatment Conditions on Observations of Student Behavior During RBWS

Source	SS	df	MS	F
Trained Students	5.70	1	5.70	.07
Trained Teachers	110.08	1	110.08	1.39
Reinforced Students	123.31	1	123.31	1.55
TS X TT	25.01	1	25.01	.32
TS X RE	14.57	1	14.57	.18
TT X RE	181.50	1	181.50	2.28
TS X TT X RE	124.67	1	124.67	1.57
Error	1271.30	16	79.47	
Total	1856.14	23		

Table 36

Mean Percent of Student On-Task Behavior
During RB and WS
By Treatment Condition

		Trained Teachers	Untrained Teachers	
$\bar{X}_{TS} = 72.9$ $SD_{TS} = 10.4$ Trained Students	Reinforced Students	73.9 76.3 72.0 $\bar{X} = 74.1$ SD = 2.2	73.1 95.4 64.8 $\bar{X} = 77.8$ SD = 15.8	
	Unreinforced Students	68.8 80.4 84.8 $\bar{X} = 78.0$ SD = 8.3	59.6 61.5 63.8 $\bar{X} = 61.6$ SD = 2.1	$\bar{X}_{re} = 75.6$ $SD_{re} = 9.3$
Untrained Students $\bar{X}_{-TS} = 73.9$ $SD_{-TS} = 7.8$	Reinforced Students	74.7 81.7 71.5 $\bar{X} = 76.0$ SD = 5.2	59.8 86.4 77.8 $\bar{X} = 74.7$ SD = 13.6	
	Unreinforced Students	73.4 81.7 66.8 $\bar{X} = 74.0$ SD = 7.5	65.1 77.7 69.6 $\bar{X} = 70.8$ SD = 6.4	$\bar{X}_{-re} = 71.2$ $SD_{-re} = 8.8$
		$\bar{X}_{TT} = 75.5$ $SD_{TT} = 5.6$	$\bar{X}_{-TT} = 71.2$ $SD_{-TT} = 11.3$	

Note: The three numbers in each cell represent the mean percent student on-task behavior per classroom.

Test Booklets

Student test booklets were examined for invalid marks made by students that would influence the machine-scored results. Data were analyzed by error type across treatments using ANOVA. Statistically significant differences were observed for the number of erasures made by observers and items left blank (not answered) by the students. Erasures were defined as the removal of any mark on the test booklet that was not a part of an answer fill-in. These marks may have been read as answers during machine scoring, so they were erased by observers. Blank items were defined as questions with no answer filled in by the student. Erasures and blank items were entered into the analyses by mean number per booklet per classroom.

Results from ANOVA on erasures (Table 37) show statistical significant main effects for trained students, $F(1, 16) = 7.51$, $p < .02$) ($\omega^2 = .227$). The mean number of erasures per trained student (Table 38) was 13.66 (SD = 5.12) and per untrained student, 35.43 (SD = 24.81) indicating that untrained students made significantly more marks that would invalidate the results from machine scoring than trained students. Due to the emphasis during student training on filling out machine scorable answer forms, a large difference would be expected in the number of erasures needed by booklets from untrained as opposed to trained classrooms. This evidence suggests that part of the student training (answer format) was successfully communicated but was apparently unrelated to student scores because of the careful way in which booklets were corrected before scoring.

Table 39 presents the results for ANOVA on the number of items left blank per student. A statistically significant main effect was found for TT ($F [1, 16] = 9.79$, $p < .01$). The estimate of ω^2 indicates that 23.6%

Table 37

Summary of Three-Way ANOVA of Treatment Conditions on the Number of Marks That Required Erasing Before Test Booklets Were Machine Scored

Source	SS	df	MS	F
Trained Students	2845.08	1	2845.08	7.51*
Trained Teachers	620.34	1	620.34	1.64
Reinforced Students	97.98	1	97.98	.26
TS X TT	739.74	1	739.74	1.95
TS X RE	34.56	1	34.56	.09
TT X RE	1.41	1	1.41	.00
TS X TT X RE	58.78	1	58.78	.16
Error	6064.70	16	379.04	
Total	10462.60	23		

*$p < .02$.

Table 38

Means and Standard Deviations for Number of Marks that Required Erasing Before Test Booklets Were Machine Scored

	$\bar{X}$	SD
+TS Classrooms	13.66	5.12
-TS Classrooms	35.43	25.81
All Classrooms	24.55	21.33

Table 39

Summary of Three-Way ANOVA of Treatment Conditions on the Number of Test Items Left Blank by the Students

Source	SS	df	MS	F
Trained Students	5.81	1	5.81	.73
Trained Teachers	77.81	1	77.81	9.79**
Reinforced Students	4.34	1	4.34	.55
TS X TT	.22	1	.22	.03
TS X RE	3.93	1	3.93	.49
TT X RE	9.83	1	9.83	1.24
TS X TT X RE	58.37	1	58.37	7.34*
Error	127,20	16	7.95	
Total	287.51	23		

*$p < .02$.

**$p < .01$.

of the variance in blank items was accounted for by training teachers. Means listed in Table 40 indicate that a difference of 3.61 items per student distinguishes trained teacher and untrained teacher conditions. These results provide some evidence as to the effectiveness of the teacher training package in communicating the importance of answering as many questions as possible.

Summary

Although not originally included as a part of the "State Refinements" workscope, and paid for largely out of other sources, this component of the project provides important information about the questions the project was designed to address. More specifically, the project was responding to concerns among SEA and LEA personnel that the results of Title I evaluation since the implementation of TIERS have seemed inconsistant with historical estimates and more variable from year to year in the same project than seemed reasonable. This component of the project identified and provided empirical evidence about three factors which may be partially responsible for the descrepencies in Title I evaluation results. Although somewhat different from the factors the project was originally designed to address (i.e., the validity of Model A and the degree to which assumptions underlying Model A are being violated in Utah schools), these factors are nevertheless important because they impact on the results of all the evaluation models and, in fact, can influence the results of any evaluation which depends on the administration of standardized tests.

The results of this component of the project present convincing evidence that the way in which a standardized test is administered and the degree to which students are motivated to do well on the test is substantially related

Table 40

Means and Standard Deviation for Number of Test Items Left Blank

	$\bar{X}$	SD
+TT Classrooms	3.86	1.86
-TT Classrooms	7.47	3.95
All Classrooms	5.67	3.54

to the scores that students receive. Since test scores are frequently interpreted as an indicator of what students know, these data indicate that other factors besides knowledge are playing a significant role in determining students' scores. Consequently, students' scores on standardized achievement tests must be interepreted cautiously and with reference to such factors as motivation and proper test administration procedures.

CHAPTER VII

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

The previous chapters of this report have described the procedures and results of a project undertaken by the Utah State Office of Education with support from the United States Department of Education to make "State Refinements to the ESEA, Title I Evaluation and Reporting System." Funding for the project was awarded under competitive bidding procedures in response to RFP.

The project was motivated primarily by a perception among LEA and SEA staff responsible for Title I programs that the results of Title I evaluations in Utah since implementation of the Title I evaluation and reporting system (TIERS) appeared to be inconsistent with previous Title I results and at times, inconsistent for a given project from year to year. These perceptions motivated the question which this project was designed to answer: Are Title I evaluation results obtained with Model A of the TIERS accurate and believable?

The project's workscope was designed to provide information about this basic question and consisted of three parts: (a) an extensive review of the literature which has considered, both empirically and philosophically, the validity of TIERS evaluation models; (b) an empirical comparison of the estimated impact of a single Title I program using both models A and B (since Model B is assumed to be the more rigorous of the TIERS models, if both A and B were properly implemented and yielded different estimates of program impact, the results could most plausibly be attributed to weaknesses in model A; (c) an investigation of the degree to which assumptions of Model A are being violated during implementation of Title I programs in Utah.

As the organized workscope was implemented, two additional components

were added. Funding for these components came from contributed resources from the Utah State Office of Education and Utah State University; small grant from the Office of Special Education; and, more efficient uses of the resources resulting from the contract with the Department of Education. These additional components included (a) an investigation of the effect of item format on students' scores on standardized achievement tests and (b) the effects on standardized achievement test performance of training teachers, training students and motivating students.

The review of previous literature about the Title I Evaluation and Reporting System was conducted primarily to establish a foundation on which the other components of the project could build. The remainder of this chapter summarizes the major findings and recommendations resulting from the project. Additional detail regarding each of the project components is provided in previous chapters.

A Comparison of TIERS Models A and B.

Findings. Not only in Utah, but nationwide, Model A is the most frequently used Title I evaluation model and the one which is most reasonable for LEAs to implement. However, the results of this study indicate that Model A, even when it is correctly implemented, appears to overestimate the impact of a Title I program. Depending upon the grade level and content area, this overestimation ranges from one-sixth of a standard deviation to more than one-half of a standard deviation. Such differences in estimated impact can hardly be considered trivial.

As important, however, are the data which suggest that either Model A or Model B can be implemented correctly but in different ways and yield very different estimates of impact for the same project. These differences within

models occur primarily because even within a given model, the pool of students on which impact is being estimated may include substantially different students.

The two different selection methods used for of Model A in this study resulted in 50% to 75% more students being included in the second selection method, although both were appropriate. The three different methods used to select students for Model B, although all technically correct, also resulted in different groups of students being considered. Especially where mobility and attrition are high, as would frequently be the case in Title I programs, the differences in these implementation methods for either Model A or Model B can result in very different estimates of program impact for a particular Title I program.

The point has been made that attrition is not a major concern with Model A because data are obtained for the students who are still in the program at the end of the year. However, it is important to note that the TIERS models result in conclusions about impact of a total program and not about the impact of a program on a given individual student. Depending on how the evaluation model is implemented, the students for whom data are available may be very different students from the students for whom data are not available. Such differences in the student population being considered in the evaluation data may lead to very different conclusions about program effectiveness. If evaluation data are to be used to make programmatic decisions about continuation and/or improvement, such differences are very important and cannot be ignored in the decision making process.

Recommendations. The findings of this component of the project place the State Office of Education in somewhat of the dilemma. Although the results indicate that Model A may be overestimating the impact of Title I programs, it

is unreasonable to expect that most school districts in the state will be in situations that would allow them to use Model B or Model C. At this point in time virtually all of the districts are opting for Model A. Rather than have some districts using Model A, some Model B, and some Model C, when there is evidence to suggest that data from the three models are not comparable it would make more sense for the state to encourage all districts to use model A. Although this may lead to some overestimation of project impact, at least data from projects will be comparable from one project to another.

In addition, the State Office should continue to work with LEAs and inlist the support of the Technical Assistance Center in working with districts as they implement Model A to ensure that the methods used for including students data in the evaluation results are explicit and well defined. This recommendation focuses on a different issue than the formal procedures and objective tests used to select students for participation in Title I programs. The results of this project suggest that even after students have been properly selected for participation in the Title I program mobility, attrition, and absenteeism, contribute to difficulties in including many students' results in TIERS. Since the results of TIERS are used to make statements about the total program, it is important for districts to have as many children who participated in the Title I program as possible included in the reporting of the evaluation results.

Degree to Which Assumptions Made by Model A are Met in Utah Title I Evaluations.

Findings. Data from this component of the project have indicated that most of the mechnical assumptions of Model A (e.g., separation of selection and pretest, testing near the empirical norming date, and using appropriate levels of the test) are adhered to reasonably well by most districts. The

most serious problem was the degree to which appropriate selection measures were used for those students who moved into the districts after the majority of the Title I students had been selected, particularly in districts where the majority of the selection occurred during the Spring. It appeared that a substantial number of students may be selected for Title I programs in ways different from the Spring selection procedures that may violate the guideline of separating selection test and the pretest. Furthermore, it was disturbing that many Title I directors and other LEA Title I personnel did not have a clear understanding of TIERS requirements for the Model they were using.

Although most districts appear to be following the mechanical assumptions of Model A reasonably well, there are other assumptions which are somewhat more subtle but in many ways more important, that appear to be violated frequently. Generally accepted test administration practices are frequently not followed and factors such as the match between the instructional emphasis and the emphasis of the standardized test used by the district are frequently sources of difficulty. Consequently, even though the TIERS models may be implemented "correctly" the results of the evaluation regarding the impact of Title I programs may be difficult to interpret.

Recommendations. The importance of following the guidelines for implementing Model A should be continually emphasized. More importantly, however, the State Office and district Title I directors should focus additional attention on following standardized test administration procedures and selecting tests which emphasize the same factors being emphasized in the Title I instruction. Regardless of which Title I evaluation Model is utilized these factors could substantially impact on the results.

The Effect of Item Format on Students' Scores from Standardized Achievement Testing

Findings. An analysis of standardized achievement tests frequently used in conjunction with Title I evaluations revealed that different tests use different types of items to assess students' mastery of the same content. In two studies where groups of Title I students were asked to answer questions about identical content using items that had been written in the various formats from frequently used standardized achievement tests, it was found that the type of format used to ask the question accounted for almost three-quarters of a standard deviation difference in students' scores.

These data suggest that what a student appears to know based on the results of a standardized achievement test may be influenced heavily by the particular format used by that test in addition to what the student really does know about the content. The reason for differences between types of format was not addressed specifically in this study but it may well be that students have greater difficulty with formats with which they are unfamiliar. Consequently, not only the match between instructional emphasis and the emphasis of the standardized tests is important, but also it is important that the students be familiar with the types of formats in which questions will be asked.

Recommendations. The State Office should continue to investigate the effect of item format on standardized testing results and emphasize to Title I personnel the importance of making sure that students are familiar with the formats that will be used in the particular standardized achievement tests used by their district. The extent to which factors (Other than the students knowledge of the content being tested) can be controlled and/or eliminated from the standardized testing, the more valid and useful results of Title I evaluations will be. Currently, it is difficult to know whether a student's low score is a result of not knowing the content being tested or, results

from the student's unfamiliarity with the particular format being used to test the content.

Effects on Standardized Test Performance of Training Teachers, Training Students and Motivating Students.

Findings. Using a true experimental design, 24 classrooms of students in Title I Schools were randomly assigned to experimental and control conditions on three factors: (a) training teachers in appropriate standardized testing procedures, (b) training students in test-taking skills, and (c) motivating students to do their best on standardized tests. The results of this study indicate that students who are tested by trained administrators and/or who are motivated to do their best on the test, do substantially better than students who are not. Coupled with the information from the projects on site visits which suggested that procedures for which standardized test administration in Title I evluations were frequently violated, data suggest that test administration techniques and student motivation variables may be confounding the results of Title I evaluations. Students who were motivated to perform well on the achievement tests scored almost one-half of a standard deviation above those who were not. Students who were administered the test by trained administrators scored almost one-quarter of a standard deviation above those who were not.

Recommendations. Continued effort needs to be made to assure that those people responsible for administering standardized achievements tests and Title I evaluations are properly trained and follow appropriate standardized testing procedures. Furthermore, efforts need to be made to motivate students to try their best on the achievement tests. The methods used in this study (i.e., paying students for scoring higher on the test than had been predicted from a

pretest) are obviously not appropriate as a standard practice. However other, more practical procedures, need to be investigated and empirically tested. If students don't care whether they do well on a test, the results of that test can hardly be used as a measure of program impact.

Summary

The results or this project raise a number of questions regarding the interpretations of Title I evaluation results. Some of those questions (e.g., the apparent inflated estimates of impact using model A) apply only to a particular model while other concerns (e.g., lack of adherence to standardized testing procedures, effects of student motivation and item format) cut across all evaluation models.

Any type of state-wide or national evaluation system is bound to be complex. The complexities in the Tiers as indicated by this research are of greater magnatude than many people have assumed and should be considered carefully in interpreting the results of Title I evaluations. The solution, obviously, is not to discard all evaluation. Evaluation is important if determinations are to be made about effectiveness. However, these results do indicate that we must be more careful in implementing the evaluation models and in interpreting the results of those models.

Furthermore, any evaluation system which utilizes standardized achievement testing to draw conclusions about how much students know in a particular content area must take into consideration the results of this research. Based on these data, it appears that a number of other factors (e.g., the way in which the test is administered, the student's level of motivation, the type of format used by the particular achievement test) are substantially related to students' score on an achievement test besides what a student actually knows. Unless these other factors can be eliminated or

controlled, it is difficult to tell how much of a student's score is a function of his or her knowledge and how much is a function of these other factors. Unless this can be determined, evaluation results, regardless of which Title I evaluation model is used, will be difficult to interpret.

This project has not provided easy or definitive answers to questions originally motivated the study. Instead, it has provided a variety of data which should make Title I administrators in both SEA's and LEA's more careful in implementing Title I evaluation and more cautious in interpreting the results. Most importantly, however, it has more clearly defined some important questions which need further investigation if the results of most Title I evaluations are to be clearly interpretable.

REFERENCES

Armor, D., Conry-Osequera, P., Cox, M., King, N., McDonnell, L., Pascal, A., Pauly, E., & Zellman, G. Analysis of the school preferred reading program in selected Los Angeles minority schools. Report prepared for the Los Angeles Unified School District, R-2007-LAVSD, Rand Corporation, Santa Monica, CA, 1976. In S. Murray, J. Arter, & B. Faddis, Title I technical issues as threats to internal validity of experimental and quasi-experimental designs: annotated bibliography. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Arter, J. A., & Estes, G. D. A model for developing local norms with a standardized achievement measure for use with local program evaluation: procedures and effects. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada, March 27-31, 1978.

Baker, R., & Williams, P. Issues related to interpolation. In Report of the committee to examine issues related to the use of the norm referenced model for Title I evaluation. Compiled by J. B. Hansen. Portland, Oregon: Northwest Regional Educational Laboratory, October, 1978.

Barcisk̀owski, R. S., & Olsen, H. Test item arrangement and adaptation level. The Journal of Psychology, 1975, 90, 87-93.

Barnes, R. E., & Ginsburg, A. L. The relevance of the RMC models for Title I policy concerns. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada, March 27-31, 1978.

Benson, B., & Crocker, L. The effects of item format and reading ability on objective test performance: A question of validity. Educational and Psychological Measurement, 1979, 39, 381-387.

Bridgeman, B. Extrapolation and interpolation in Model A1 Title I evaluation. In Report of the committee to examine issues related to the use of the norm referenced model for Title I evaluation. Compiled by J. B. Hansen. Portland, Oregon: Northwest Regional Educational Laboratory, October, 1978.

Burton, B. Model A1 and the regression effect. Memo to TAC staff, May 8, 1978. In S. Murray, J. Arter, & B. Faddis, Title I technical issues as threats to internal validity of experimental and quasi-experimental designs: annotated bibliography. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Carcelli, L., Taylor, C., & White, K. The effect of item format on phonics subtest scores of standardized reading achievement tests. Paper presented at the Annual Meeting of the American Educational Research Association, 1981, Los Angeles, California.

Catts, R. How many options should a multiple choice question have? (a) 2. (b) 3. (c) 4. At a glance research report. Sydney, Australia: New South Wales Department of Education, 1978. (ERIC Document Reproduction Service No. ED 173 354)

Cochran, W. G. The use of covariance in observational studies. Applied Statistics, 1969, 18, 270-275. In G. Echternacht & S. Swinton, Getting straight: everything you always wanted to know about the Title I regression model and curvilinearity. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April 8-12, 1979.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPortland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. Equality of educational opportunity. U.S. DHEW, Office of Education, Washington D.C.: Government Printing Office, 1966. In S. Murray, J. Arter, & B. Faddis, Title I technical issues as threats to internal validity of experimental and quasi-experimental designs. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Conklin, J. The effects of date of testing and method of interpolation on the use of standardized test scores in the evaluation of large-scale educational programs. Paper presented at the annual convention of the American Educational Research Association, San Francisco, April, 1979.

Crane, L. R., & Cech, J. Title I evaluation models A1 and B1: an empirical comparison. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Cronbach, L. J. Test validation. In Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Crowder, C. R., & Gallas, E. J. Relation of out-of-level testing to ceiling and floor effects on third and fifth grade students. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada, March 27-31, 1978.

Dambrot, F. Test item order and academic ability, or should you shuffle the test item deck? Teaching of Psychology, 1980, 7, 94-96.

David, J. L., & Pelavin, S. H. Evaluating compensatory education: Over what period of time should achievement be measured? Journal of Educational Measurement, 1978, 15(2), 91-99.

DeVito, P. J., & Long, J. V. The effects of spring-spring vs. fall-spring testing upon the evaluation of compensatory education programs. Paper presented at the annual convention of the American Educational Research Association, New York City, April, 1977.

Doherty, W. Restandardization study. System Development Corporation (undated). In S. Murray, J. Arter, & B. Faddis. Title I technical issues as threats to internal validity of experimental and quasi-experimental designs. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Echternacht, G. Model C is feasible for ESEA Title I evaluation. Paper presented at the annual meeting of the American Educational Research Association, Boston, Mass., April 10, 1980.

Echternacht, G., & Swinton, S. Getting straight: Everything you always wanted to know about the Title I regression model and curvilinearity. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April 8-12, 1979.

Estes, G. D., & Anderson, J. I. Observed treatment effects with special regression evaluation models in groups with no treatment. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada, March 27-31, 1978.

Estes, W. K., & DePolito, F. J. Independent variation of information storage and retrieval processes in paired associate learning. Journal of Experimental Psychology, 1967, 75, 18-26.

Faddis, B. J., Arter, J. A., & Zwertchek, A. An empirical comparison of ESEA Title I evaluation models A and B. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 8-12, 1979.

Faddis, B. J., & Estes, G. D. Fall-to-spring vs. fall-to-fall evaluation of a large Title I program with a comparison group design. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Fagan, B. M., & Horst, D. P. Selecting a norm-referenced test. Mountain View, CA: RMC Research Corporation, 1978. In R. T. Johnson & W. P. Thomas, User experiences in implementing the RMC Title I evaluation models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Fish, O. W. An analysis of the evaluation data when ESEA, Title I evaluation models A1 and A2 are empirically field tested simultaneously. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California, April 8-12, 1979.

Fishbein, R. L. The use of non-normed tests in the ESEA Title I evaluation and reporting system: Some technical and policy issues. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada, March 27-31, 1978.

Frisbie, D. A. Multiple choice versus true-false: A comparison of reliabilities and concurrent validities. Journal of Educational Measurement, 1973, 10, 297-304.

Gabriel, B. R., Stenner, A. J., & Troy, J. B. An empirical examination of three models for estimating the effects of no-treatment. Paper presented at the American Educational Research Association Convention, New York City, April 6, 1977.

Gerow, J. R. Performance on achievement tests as a function of the order of item difficulty. Teaching of Psychology, 1980, 7, 93-94.

Glass, G. Memo to: unknown, interested parties. In Report of the committee to examine issues related to the use of the norm referenced model for Title I evaluation. Compiled by J. B. Hansen. Portland, Oregon: Northwest Regional Educational Laboratory, October, 1978.

Glass, G. V. Integrating findings: The meta-analysis of research. Review of Research in Education, 1977, 5, 351-379.

Glass, G. V, & Hakstian, A. R. Measures of association in comparative experiments: Their development and interpretation. American Educational Research Journal, 1969, 6(3), 403-414.

Goldman, J., & Crane, L. R. Title I model B adjustment procedures: which to use and when. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 7-11, 1980.

Grier, L. B. The number of alternatives for optimum test reliability. Journal of Evaluation Measurement, 1975, 12, 109-112.

Halpin, G., & Halpin, G. Retention in an actual classroom setting as a function of type and complexity of tests. (ERIC Document Reproduction Service No. ED 183 622)

Hardy, R. A comparison of Model A and Model C: Results of first year implementation in Florida. ETS, Evanston, Ill. Personal communication, January, 1979. In S. Murray, J. Arter, & B. Faddis, Title I technical issues as threats to internal validity of experimental and quasi-experimental designs (Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979) and G. Echternacht, Model C is feasible for ESEA Title I evaluation (Paper presented at the annual meeting of the American Educational Research Association, Boston, Mass., April 10, 1980).

Hays, W. L. Statistics for the social sciences. New York: Holt, Rinehart and Winston, 1973.

Hiscox, S. B., & Owen, T. R. Behind the basic assumption of Model A. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada, March 27-31, 1978.

Holliday, W. G., & Partridge, L. A. Differential sequencing effects of test items on children. Journal of Research in Science Teaching, 1979, 16, 407-411.

Horst, D. P., & Wood, C. T. Collecting achievement test data. Mountain View, CA: RMC Research Corporation, 1978. In R. T. Johnson & W. P. Thomas, User experiences in implementing the RMC Title I evaluation models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

House, G. D. A comparison of Title I achievement results obtained under USOE models A1, C1 and a mixed model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California, April 8-12, 1979.

Johnson, D. D., Pittleman, S., Ackwenker, J., & Perry, J. On bias diagnosis of reading difficulties - IV (Technical Report No. 464). Madison: Wisconsin Research and Development Center for Individualized Schooling, 197?. (ERIC Document and Reproduction Service No. ED 173 355)

Johnson, R. T., & Thomas, W. P. User experiences in implementing the RMC Title I evaluation models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

The Joint Dissemination Review Panel. IDEABOOK. Washington, D.C.: Superintendent of Documents, U.S. Government Printing Office, October, 1977.

Kaskowitz, D. H., & Norwood, C. R. A study of the norm referenced procedure for evaluating project effectiveness as applied in the evaluation of project information packages. Stanford Research Institute, Research Memorandum URU-3556, Menlo Park, CA, January, 197_). In S. Murray, J. Arter, & B. Faddis, Title I technical issues as threats to internal validity of experimental and quasi-experimental designs: Annotated bibliography. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Kenny, D. A. A quasi-experimental approach to assessing treatment effects in the non-equivalent control group design. Psychological Bulletin, 1975, 82, 345-362. In J. Goldman & L. R. Crane, Title I Model B adjustment procedures: which to use and when. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 7-11, 1980.

Kierscht, M. S., & Vietze, P. M. Test stimuli: Representational level with middle class and head start children. Psychology in the Schools, 1975, 12, 309-312.

Kintsch, W. Recognition and free recall of organized lists. Journal of Experimental Psychology, 1968, 78, 481-487.

Kumar, V. K., Rabinsky, L., & Pandley, T. N. Test mode, test instructions, and retention. Contemporary Educational Psychology, 1979, 4, 211-218.

Lai-Min, P. S., & Coffman, W. E. A study of "I don't know" response in multiple-choice tests. Iowa City, Iowa: Iowa Testing Programs, University of Iowa, 1974. (ERIC Document Reproduction Service No. ED 141 371)

Linn, R. L. The validity of inferences based on the proposed Title I evaluation models. Paper presented as part of a symposium at the annual meeting of the American Educational Research Association, Toronto, Canada, March 27-31, 1978.

Linn, R. L., & Werts, C. E. Analysis of implications of the choice of a structural model in the non-equivalent control group design. Psychological Bulletin, 1977, 84, 229-234.

Loftus, G. R. Comparison of recognition and recall in a continuous memory task. Journal of Experimental Psychology, 1971, 91, 220-226.

Long, J., Horwitz, S., & DeVito, P. An empirical investigation of the ESEA Title I evaluation systems' proposed variance estimation procedures for use with criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Long, J. V., Schaffran, J. A., & Kellogg, T. M. Effects of out-of-level survey testing on reading achievement scores of Title I, ESEA students. Journal of Educational Measurement, 1977, 14, 203-213.

Mayeske, G. W., & Beaton, A., Jr. Special studies of our nation's students. Washington, D.C.: Government Printing Office, 1975. In S. Murray, J. Arter, & B. Faddis, Title I technical issues as threats to internal validity of experimental and quasi-experimental designs. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Millman, J., & Setijadi. A comparison of American and Indonesian students on three types of test items. Journal of Educational Research, 1966, 59, 273-275.

Mishra, S. P. Influence of the examiner's ethnic attitudes on intelligence test scores. Psychology in the Schools, 1980, 17, 177-122.

Murray, S., Arter, J., & Faddis, B. Title I technical issues as threats to internal validity of experimental and quasi-experimental designs: Annotated bibliography. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Noggle, N. L. Alternative norms for model A1. NWREL/TAC, 1977. Draft. In S. Murray, J. Arter, & B. Faddis, Title I technical issues as threats to internal validity of experimental and quasi-experimental designs: Annotated bibliography. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Ozenne, A. The effect of vertical scaling imprecision in the estimation of Title I project gains. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978. In S. Murray, J. Arter, & B. Faddis, Title I technical issues as threats to internal validity of experimental and quasi-experimental designs. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Piersel, W. C., Brody, G. H., & Kratochwill, T. R. Further examination of motivational influences on disadvantaged minority group children's intelligence performance. Child Development, 1977, 48, 1142-1145.

Poage, M., & Poage, E. G. Is one picture worth a thousand words? Arithmetic Teacher, 1977, 24, 408-414.

Porter, A. C., Schmidt, W. H., Floden, R. E., & Freeman, D. J. Practical significance in program evaluation. American Educational Research Journal, 1978, 15, 529-539. In S. Murray, J. Arter, & B. Faddis, Title I technical issues as threats to internal validity of experimental and quasi-experimental designs. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Powell, G., Schmidt, J., & Raffeld, P. The equipercentile assumption as a pseudo-control group estimate of gain. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April 8-12, 1979.

Powers, S., & Gallas, E. J. Implications of out-of-level testing for ESEA Title I students. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada, March 27-31, 1978.

Report of the committee to examine issues related to the use of the norm referenced model for Title I evaluation. Compiled by J. B. Hansen. Portland, Oregon: Northwest Regional Educational Laboratory, October, 1978.

Roberts, S. J. Test floor and ceiling effects. Mountain View, CA: RMC Research Corporation, 1978.

Roid, G., & Haladyna, T. A review of item writing methods for criterion-referenced tests in the cognitive domain. Oklahoma City, Oklahoma: Paper presented at the Annual Meeting of the Military Testing Association, 1978. (ERIC Document Reproduction Service No. ED 178 562)

Slaughter, H. B., & Gallas, E. J. Will out-of-level norm-referenced testing improve the selection of program participants and the diagnosis of reading comprehension in ESEA Title I programs? Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada, March 27-31, 1978.

Stonehill, R. M., & English, J. J. Measurement concerns in Title I evaluation. Paper presented at the Large School Systems' Invitational Conference on Measurement and Evaluation held in Alexandria, Virginia, May 7, 1979.

Storlie, T. R., Rice, W., Harvey, P., & Crane, L. R. An empirical comparison of Title I NCE gains estimated with model A1 and with model A2. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Tallmadge, G. K., & Horst, D. P. The use of different achievement tests in the ESEA Title I evaluation system. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada, March 27-31, 1978.

Tallmadge, G. K., & Horst, D. P. Using the data from state and local ESEA Title I reports. Paper presented at the annual meeting of the American Educational Research Association, New York City, April 4-8, 1977.

Tallmadge, G. K., & Roberts, A. O. H. Factors that influence test results. Mt. View, CA: RMC Research Corporation, 1978. In R. T. Johnson & W. P. Thomas, User experiences in implementing the RMC Title I evaluation models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Tallmadge, G. K., & Wood, C. T. Comparability of gains from the three models in the Title I evaluation system. Mountain View, CA: RMC Research Corporation, 1980.

Tallmadge, G. K., & Wood, C. T. User's guide: ESEA Title I evaluation and reporting system. Prepared for U.S. Dept. of Health, Education, and Welfare. Mountain View, CA: RMC Research Corporation, October, 1976.

Taylor, C. Personal communication. February 20, 1981.

Towle, N. J., & Merrill, P. F. Effects of anxiety type and item difficulty sequencing on mathematics test performance. Journal of Educational Measurement, 1975, 112, 241-249.

Van Hove, E., Coleman, J. S., Rabben, K., & Karweit, N. Schools' performance: New York, Los Angeles, Chicago, Philadelphia, Detroit, Baltimore. Unpublished manuscript, Baltimore, Md., October, 1970. Reported in R. L. Linn, The validity of inferences based on the proposed Title I evaluation models. Paper presented as part of a symposium at the annual meeting of the American Educational Research Association, Toronto, Canada, March 27-31, 1978.

Weber, M. B. The effect of choice format on internal consistency. Emory University. (ERIC Document Reproduction Service No. ED 161 940)

Weiten, W. Relative effectiveness of single and double multiple-choice questions in educational measurement. New York, N.Y.: Paper presented at the Annual Meeting of the American Psychological Association, 1979. (ERIC Document Reproduction Service No. ED 185 097)

Williams, R. L., Davis, W., Anderson, P., & Favor, K. Test format as a form of bias for black students. Journal of Non-white Concerns in Personnel and Guidance, 1978, 6, 141-147.

Yap, K. O., Estes, G. D., & Hansen, J. B. Effects of data analysis methods and selection procedures in regression models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Yap, K. Y. Can selection tests be used as pretests? Paper presented as part of a symposium at the annual meeting of the American Educational Research Association, Toronto, Canada, March 27-31, 1978.

Appendix

Letters to District Title I Directors
Explaining Purpose of Project

UTAH STATE UNIVERSITY · LOGAN, UTAH 84322

801-750-1981

UNIVERSITY AFFILIATED
EXCEPTIONAL CHILD CENTER
UMC 68

March 24, 1980

Dear

Recently you received the attached letter from Kent Worthington explaining that the Utah State Office of Education had received a contract from the United States Office of Education to investigate the evaluation models required by the new Title I Evaluation and Reporting System (TIERS). Staff from the Psychology Department at Utah State University have been asked to assist the State Office of Education in collecting data regarding the effectiveness and applicability of Model A. In particular we will be studying Model A's underlying assumptions about testing levels, dates and procedures, and whether they are relevant to the real needs of the school situation.

Hopefully, the investigation of these assumptions will help LEA's and SEA's to:

a) make better informed decisions about the selection of a local evaluation model for Title I programs;

b) better interpret the results from evaluations using Model A; and

c) avoid the violation of Title I assumptions.

To help achieve these objectives we would like to interview LEA personnel in thirteen Utah School Districts about the procedures of test administration, the rationale for test selection, and any type of problems they may have had in implementing the models. We would also like to observe some of the Title I testing in each district to verify and expand the data collected during the interviews.

The interview and observation data would be collected by trained Utah State University graduate students during the time you normally administer the standardized test in conjunction with Title I. The interviewing and observation procedures have been designed to be as unobtrusive as possible. During the coming week we will contact you by phone to answer any questions you have about the project and explain in more delail what we would like to accomplish during the visit. At that time we would also appreciate your assistance in identifying which school or schools would be best to visit.

Although particaption in the project is voluntary, your cooperation will greatly enhance the effectiveness of the study. Should you have questions or concerns you would like to discuss before we are able to talk with you on the phone, feel free to contact Kent Worthington (801 533-6092), or myself (toll free 800 622-5420). I look forward to talking with you more in the near future.

Sincerely;

Karl R. White, PhD
Director Planning and Evaluation,
Exceptional Child Center and
Assistant Professor of Psychology

250 EAST 500 SOUTH STREET · SALT LAKE CITY, UTAH 84111 · TELEPHONE (801) 533-5431

UTAH STATE OFFICE OF EDUCATION

WALTER D TALBOT
STATE SUPERINTENDENT OF PUBLIC INSTRUCTION

March 18, 1980

Dear

As you know, all states are now required to use an Evaluation Model from the Title I Evaluation and Reporting System (TIERS) to evaluate the impact of Title I projects within the state. Your district is one of 13 districts within the state which will be reporting Title I evaluation data to the Feds during 1979-80 as a part of the state's three year evaluation plan.

Most districts within Utah have chosen to use Model A to conduct their Title I evaluation. The results obtained from most school districts have been positive and demonstrate student gains beyond achievement from the regular instructional program. We have observed considerable variance in student gains among grades, schools, and districts. Also, the amount of gain varies among different evaluation models, e.g., A, B, and C, and pre-post testing time intervals, e.g., fall to spring, and spring to spring. In order to determine the reasons for the variances observed and to increase confidence in achievement data, evaluation workshops and individual consultation sessions have been conducted in most district. State Office and Northwest REL-TAC personnel have invested considerable time in these activities in recent months.

The U.S. Education Department has been concerned. State refinement contracts have been authorized to study such matters in greater depth. Last summer a contract was granted to the Utah State Office of Education; it is being accomplished by Karl White and other evaluators at Utah State University. They are studying the effects of Model B in depth in one district and will be doing field work in selected districts.

As one part of this project, we would like to collect additional data about the implementation of Model A in each of the districts which will be reporting evaluation results this year. Personnel from Utah State University will be assisting us in this component of the project and will be contacting your district shortly to coordinate times and procedures for this data collection (in some cases, preliminary contact has already been made). Data collection in each district will consist of a limited amount of observational data collected during

AVARD A. RIGBY, Administrator
Division of Program Administration
Telephone (801) 533-5061

Adult Education and Community Services
Dissemination and Educational Development
Telephone (801) 533 5061
Compensatory and Bilingual Education
(801) 533-6092
Special Education (801) 533-5982

Title I Spring Testing and brief interviews conducted with a number of the district staff. We have tried to plan these activities so that minimal disruption of your regular activities will occur. Data collected during these visits will be used to draw conclusions about the technical adequacy of Model A and will not be used to make funding decisions or draw conclusions about the "quality" of a particular district's Title I program. Although your participation in this project is, of course, voluntary, full participation will contribute greatly to the success of the project.

If you have additional questions, you can refer them to us at the Utah State Office of Education or to Dr. Karl White, Director of Planning and Evaluation, Exceptional Child Center, Utah State University (toll free phone 800-662-5420). In addition, we would be happy to send you a full copy of the funding proposal which describes the project in more detail. We look forward to working with you in this important project.

Sincerely yours,

Kent L. Worthington, Coordinator
Title I, ESEA

Jay K. Donaldson, Specialist
Title I, ESEA

dd

Appendix 2

Letter to Principals of Schools
Visited During Project

UTAH STATE UNIVERSITY, LOGAN, UTAH 84322

801-750-1981

UNIVERSITY AFFILIATED
EXCEPTIONAL CHILD CENTER
UMC 68

March 26, 1980

Dear

The Utah State Office of Education has received a contract from the United States Office of Education to conduct a study of the evaluation models which the federal government now requires all states to use in the Title I evaluation and Reporting System (TIERS). This project is under the direction of the State Title I Director, Dr. Kent Worthington who is being assisted by selected staff from the Psychology Department at Utah State University.

Recently, we corresponded with your district superintendent and spoke with your district Title I director and they agreed to have your district participate in the project. As noted in the attached letter from Dr. Worthington, the basic purpose of the project is to evaluate the effectiveness and applicability of Title I evaluation Model A for collecting data about the state's Title I programs. Data from the project will not be used to make funding decisions or statements or worth about an individual districts Title I program.

As a part of the project we would like to visit your school during the time you are conducting the post testing for your Title I program. During this time we want to observe students' reactions to the testing and interview school faculty about their preceptions of the strenghts and weaknesses of the Model A evaluation. These data would be collected by trained graduate students from Utah State University. Data collection procedures have been designed to be as unobtrusive as possible and will require very little direct time from any individual member of your staff.

It is our understanding that you will be doing your post testing during the week of . Shortly after you receive this letter, we will contact you by phone to answer any questions you have and, if you agree to participate, work out the details of our visit.

Should you have questions before we contact you, please feel free to call Kent Worthington (801 533-6092) or myself (toll free 800 662-5420). I look forward to talking with you more in the near future.

Sincerely,

Karl White
Director
Planning & Evaluation
Exceptional Child Center and
Assistant Professor of Psychology

Appendix 3

Memorandum Provided to Principals for Informing Teachers About the Project

UTAH STATE UNIVERSITY
University Affiliated
Exceptional Child Center

MEMORANDUM

To:

From:

Subject: USOE Study Concerning Title I Evaluation Models

Date:

The Utah State Office of Education has received a contract from the United States Office of Education to investigate the effectiveness and applicability of the Title I Evaluation and Reporting System (TIERS) which the federal government now requires all participating Title I programs to use in evaluating their projects. As a part of the study, project staff will be visiting our school on During their visit they will be observing students who are taking tests and talking briefly with some of us about the testing procedures and our reactions to the present system of evaluating our Title I program.

Observation of the testing should not disrupt your normal operation at all, but I wanted you to be aware of the study so you would not be surprised by the presence of an unfamiliar person. At 8:00 on the day of their visit, project staff will be available in room # for anyone who has questions or would like additional information about the project. In addition I will be contacting some of you to arrange for a time (approximately 15 minutes) that you could visit with one of the project staff about some of your preceptions of the currently used Title I evaluation system. Should you have any questions, please feel free to contact me.

Appendix 4

Interview Guide Sheet for Collecting Data From
LEA Personnel Regarding Implementation
of TIERS

INTERVIEW
TITLE I PERSONNEL

DISTRICT _____ POSITION _____

SCHOOL _____ GRADE _____

DATE _____ INTERVIEWER _____

1. Student's reaction to testing:
   a) How do students feel about the testing (positive, negative, apathetic)?
   b) Do they understand the purpose of the tests?
   c) How do the students usually behave during testing?
   d) Do they try to do their best on the tests?

2. District personnel reaction to testing:
   a) Do you think the testing is worthwhile--worth the time and effort it takes?
   b) Do you use pre/post subtest for purposes other than to compare gains? Specifically how?
   c) Does anyone individually discuss with students/parents the results of the Title I testing?

3. Selection of students:
   a) What is the selection process? Per cent of out of level?
   b) Does the selection process work? Do you think the correct students are being selected?
   c) Separation of pre and selection test for <u>all</u> students.
   d) How are new move-ins selected? What percentage of total students are new move-ins?

4. Test administration:

a) Do administration dates match empirical norm dates (pre and post)? (When did you administer your pre test?). If not, do districts do any extrapolation?

b) What types of things are done to prepare students for testing? Teacher preparation?

c) How did you select particular test and form/level used in selection, pre and post? Per cent of out of level?

d) When and how are make-ups done? Estimate percentage of students who miss original testing and 1) take make-up and 2) never get make-up.

e) Who is responsible for turning data from Title I testing into USOE?

1) Is reporting format any good (strengths and weaknesses)?

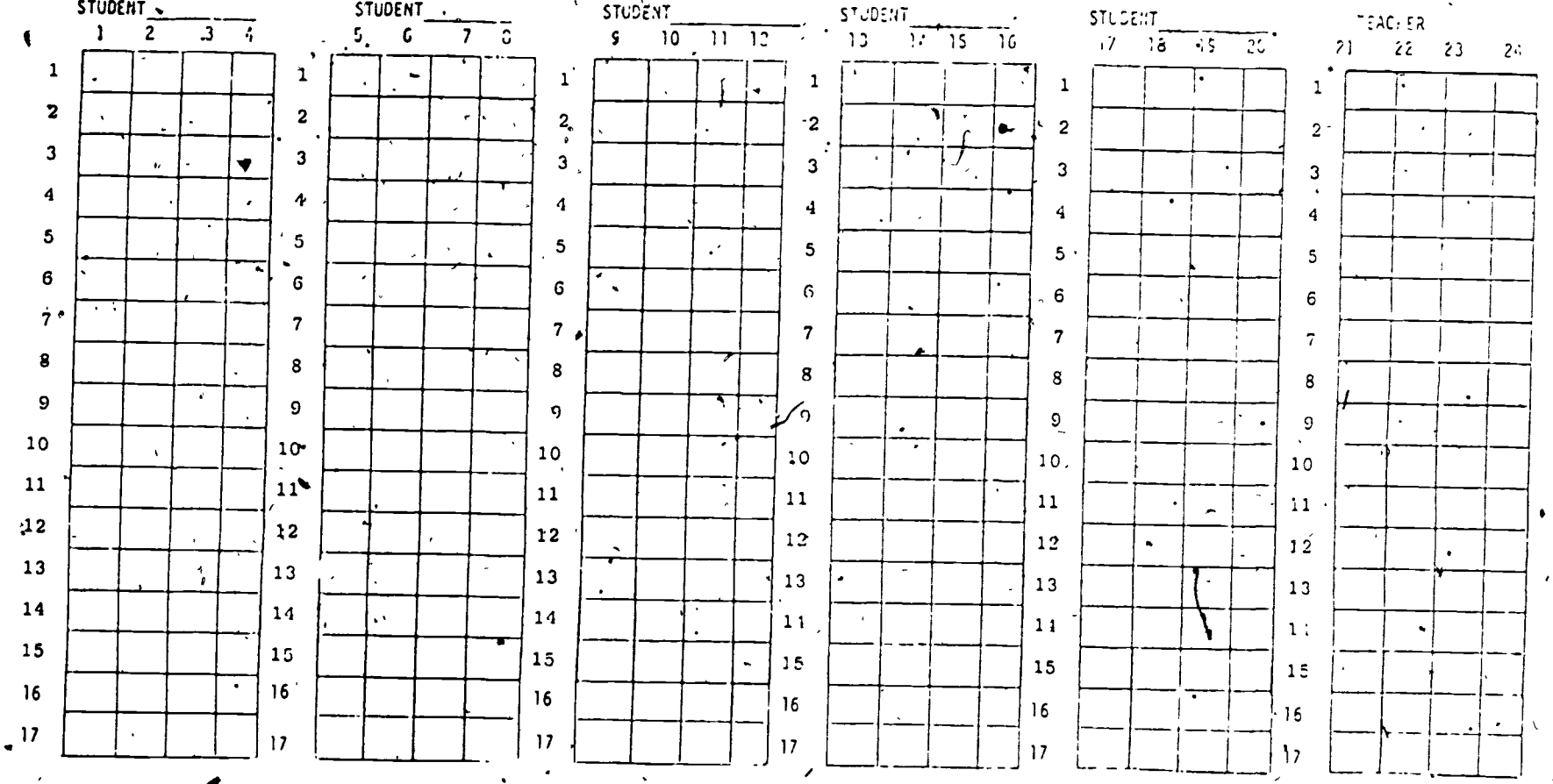
2) What checks are made to assure accuracy?

Appendix 5

Data Collection Form and Definitions of
On-Task/Off-Task Behavior for
Classroom Observation

OBSERVATION FOR TITLE I TESTING

n of Class _____ SCHOOL _____ CLASSROOM _____ DISTRICT _____ DATE _____ OBSERVER _____



TOTAL ONTASK ____ TOTAL ONTASK ____ TOTAL ONTASK ____ TOTAL ONTASK ____ TOTAL ONTASK ____ TOTAL ONTASK ____

% ONTASK____ % ONTASK____ % ONTASK____ % ONTASK____ % ONTASK____ % ONTASK____

NOTES:

INTERVALS: 3 second observation
2 second record

CODE:
- [1] Ontask (for entire interval)
- [-] Off task (for any part of interval)
- [+] Teacher contact (verbal or physical interaction with student alone)
- [*] Student is finished with test
- [ ] Beginning of timed test
- [ ] End of timed test
- [/] No record made (Explain in NOTES section)

Directions: Record 4 intervals on one student before observing next student. Observe 5 students and one teacher for a total of 24 intervals before repeating sequence.

TOTAL ONTASK = Number of "1" recorded for one student
% ONTASK = Number of "1" ÷ 32

STUDENT
ON-TASK BEHAVIOR
DURING TEST TAKING

Behavior	Teacher Directed		Student Directed	
	Example	Nonexample	Example	Nonexample
Raising hand	while looking at the teacher: -during direction giving -during test taking	while looking at: -another student -the test	while looking at the teacher -during test taking	while looking at: -another student -the test
Asking questions	concerning: -directions	concerning: -answer to questions -drinking water -using bathroom -broken pencil lead -eraser	concerning: -directions	concerning: -answers to questions -drinking water -using bathroom -broken pencil lead -eraser
Looking	at: -teacher -test paper when so directed -board when so directed -fingers if counting	at: -another student -another's test -observers -desk (inside, outside) -toys -fingers if not counting -wrong test page -clothes	at: -test paper -teacher when hand raised -prepared material for early finishers	at: -another student -another's test -observers -teacher unless hand raise -toys -fingers if not counting -desk (inside, outside) -wrong test page -clothes
Talking (audible)	to: -teacher when called upon	to: -another student -self (if words audible) -teacher when not called upon	to: -teacher when called upon	to: -another student -self (if words audible) -teacher when not called upon
Body movement	picking up pencil from floor scooting chair or desk less than 10 inches scratching body writing answers to test questions when directed playing with clothes.	writing on desk standing up (body leaves seat) hands on another's desk hands on another student or teacher writing answers to questions when not directed throwing anything kicking another's chair or desk leaning back on chair tapping pencil	picking up pencil from floor scooting chair or desk less than 10 inches scratching body writing answers to test questions when directed playing with clothes	writing on desk standing up hands on another's desk hands on another student or teacher writing answers to questions when not directed throwing anything kicking another's chair or desk leaning back on chair tapping pencil

Definitions

Teacher Directed:
Teacher gives directions
Teacher reads the question for each item.

Student Directed:
Students work at their own pace throughout test
Test is usually timed
Begins when teacher says "Ready? Go!"
Ends when teacher says "Stop! Close your booklet---."

TEACHER BEHAVIOR

DEFINITION OF ON-TASK BEHAVIOR

Behavior	Example	Nonexample
Talking	To class or individual student: - explain the directions or answer format - answer questions about directions - give directions - read questions - praise listening or working To aide: - but only to alert to non-attending student - if students are on incorrect item or have a broken pencil lead	To class or individual student: - if students are on incorrect item or have a broken pencil lead - to explain answer - to help formulate an answer - to threaten, criticize, or reprimand - to repeat questions that are to be given only once To class during TT. To aide: - except to alert To another teacher To communication system
Moving	Standing in front of room Pointing to nonattenders Providing a pencil	Standing with back to any student or where faces cannot be seen. Sitting Lying down.
Looking	At individuals in the class - after reading each sentence in the directions - after reading each question At clock At aide to alert to nonattending student	At: - another teacher - textbook - lesson plans - classroom equipment - magazine/book - manual (only during TT)

Appendix 6

Quality of Test Administration Checklist

SCHOOL _____ GRADE _____

DISTRICT _____ DATE _____

SESSION _____ OBSERVER _____

DID THE TEACHER DO THESE BEFORE ADMINISTERING THE TEST?

Class Environment

_____ 1. Arrange the students' desks so they are not touching.

_____ 2. Position the desks to face the same direction (every booklet and student's face can be seen from the front of the room).

_____ 3. Assure that the room is comfortable (temperature, light, noise).

_____ 4. Post a "Testing, Do Not Disturb" sign on door.

_____ 5. Have a visible supply of pencils.

_____ 6. Have a visible clock or watch with a minute hand.

_____ 7. Create a generally positive climate that promotes good work habits and is without pressure or tension.

_____ 8. Seat the most frequently nonattending students in the front.

Student Preparation

_____ 1. Provide an opportunity for using the bathroom, drinking water, and sharpening pencil.

_____ 2. Provide all students with a pencil and an eraser.

_____ 3. Ask students to remove nontesting material from desks if appropriate.

_____ 4. Explain the reason for the test (to use the information to help teach students).

_____ 5. Obtain the attention of the entire class for 1 minute prior to directions (all students watching teacher).

_____ 6. Pass out test booklets in less than 2 minutes and in an orderly and efficient manner.

_____ 7. Verbally reward attentive behavior.

Reminders

_____ 1. Not to leave their seats but to raise a hand if something is needed.

_____ 2. What to do if they finish before time is up (TT only).

_____ 3. To check their work if they finish before the time is up (to see if every question is answered only once).

_____ 4. That some of the items will be more difficult than their daily work.

_____ 5. To skip an item that they don't know and go on to the next one.

Positive Atmosphere

_____ 1. Praise individual students for appropriate behavior.
_____ 2. Praise class for listening and working.
_____ 3. Smile frequently.
_____ 4. Make less than two reprimands, threats, or criticisms during the subtest.
_____ 5. Speak with a gentle, but firm voice.
_____ 6. Use physical touch to prompt and reward on task behavior.
_____ 7. Start the test directions within several minutes of sitting down so that students did not become restless with preparation activities.
_____ 8. Quickly supply a student with pencil or eraser when needed.
_____ 9. Stand near front of room where all students can see easily.

Reading Directions

_____ 1. Look at class between sentences.
_____ 2. Survey the class to check if directions were followed (i.e. "Put your finger on the sample," "fill in the circle," "write your name," "turn to page 12").
_____ 3. Alert the aide to nonattenders.
_____ 4. Proceed to next direction only after all students are ready.
_____ 5. Supplement printed directions with verbal and visual explanations when students do not understand the procedure.
_____ 6. Change wording of directions to a vocabulary the students are familiar with (i.e. "circle" instead of "oval" or "box" instead of "frame".

Reading Test Items

_____ 1. Look up after each question and glance around room.
_____ 2. Follow the exact wording of questions as stated in the manual. (Never define or explain words or illustrate procedures.)
_____ 3. Allow approximately 10 seconds between items.
_____ 4. Never repeat a question unless the directions specify to do so.
_____ 5. Alert aide to nonattenders or to students with raised hands.

Timed Tests

_____ 1. Set clock for correct time requirement.
_____ 2. Watch students during entire test to detect speeding, slow answering, day dreaming and cheating.
_____ 3. Alert aide to nonattenders or to students with raised hands.

End of Test

_____ 1. Praise students for working hard.
_____ 2. Collect booklets in a directed manner.
_____ 3. Provide a directed, stand-up, rest period.

Appendix 7

Interview Guide Sheet for Teachers to Prioritize Curriculum Areas

TEACHER RATING OF ITEMS FROM STANDARDIZED TESTS

DISTRICT _____ POSITION _____

SCHOOL _____ GRADE _____

DATE _____ INTERVIEWER _____

Rank by Importance	% of Teaching Time	Category	Rank by Importance
_____	_____	A. Phonics	
		1. Consonant sounds (blends)	_____
		2. Vowel sounds (long and short)	_____
		3. Consonant digraphs (ch, wh, sh, th)	_____
		4. Vowel digraphs/diphthongs (ai, ay, ea, ee, oi, ow, etc.)	_____
		5. Controlled vowels (al, or, aw, etc.)	_____
		6. Variant vowels (said, was, etc.)	_____
		7. Other _____	_____
_____	_____	B. Vocabulary	
		1. Word meaning	_____
		2. Contextual clues	_____
		3. Analogies	_____
		4. Sight Vocabulary	_____
		5. Other _____	_____
_____	_____	C. Literal Comprehension	
_____	_____	D. Inferential Comprehension	
_____	_____	E. Structural Analysis	
		1. Root words	_____
		2. Syllabication	_____
		3. Affixes	_____
		4. Compound words	_____
		5. Contractions	_____
		6. Other _____	_____
_____	_____	F. Other:	
_____	_____	_____	
_____	_____	_____	
_____	_____	_____	
_____	_____	_____	

Total 100%

Appendix 8

Computing Student/Teacher Ratios for Title I Programs

HOW TO FIGURE STUDENT / TEACHER RATIO

(With some good luck)

You may have found, as we have, that determining the student/teacher ratio of a Title I program can be a trying and frustrating experience. The current instructions are vague and do not address the special needs of unique treatment activities. It is, of course, impossible to develop instructions that will account for every contigency. These instructions were developed to better clarify the student/teacher ratio computation procedure for a greater variety of programs. Unfortunately, the broader applicability has necessitated a greater complexity. We hope it is comprehensible.

The computation procedure will be explained by a series of directions, and illustrated by hypothetical examples. These examples have been arranged in the appendix in a grid pattern. Each direction will have a letter and/or number associated with it that refers to particular space in the grid. The grid is a suggested format for compiling ten types of data. How to do so and what to do with it will be explained.

We have indentified six basic teaching modes. A small group with more than one grade level-row A, a small group that also serve non-target sudents-row B, a large group-row C, individualized instruction-row D, and peer group tutoring-row E.

COLUMN 1 - MODE OF ACTIVITY

The mode of activity is the type of Title I treatment situation in which the Title I target children of a particular grade participate. This could be a small group activity. If there is more than one small group, each group must be entered individually (1A and 1B = 5G). Other modes

could be large groups (1C = LG), or individualized instruction (1D and 1F = II). Again individualized instruction is entered twice, but in this instance, because there are two different instructors (10 D and 10 F = Aide and PO).

The most common variant encountered is supervised peer tutoring. This variant should be entered as a small group with an explanation (1E =5G). Peer tutors are not paid Title I employees and cannot therefore be considered instructors. Unsupervised peer tutoring should not be entered at all.

COLUMN 2 - TARGET-GRADE LEVEL STUDENTS

Column 2 refers to the number of students from a particular grade that are involved in each separate activity listed in column 1. Some children are probably involved in more than one Title I treatment activity. For instance, one child may be in a small group at one time, a large group at another, and individually tutored at still another time. This child should be entered three times, once for each activity. It is obvious that the total number of students involved in the various activities in a certain grade level will, generally, far outnumber the total number of target students for that grade. This occurs because the students may be counted more than once.

A problem arised in those programs in which more than one grade level is involved in the same Title I treatment activity. Enter only the number of students for the grade of immediate concern in column 2. Row A shows a situation in which there is a small group of seven students of mixed grades (A6 = 7), but only three of which are in the second grade (A2 = 3).

A related problem arises when non-target children are sometimes included in Title I treatment groups. This case occurs in row B. In this case, enter only the number of target children being served in column 2 (2B = 8).

COLUMN 3 - LENGTH OF ACTIVITY.

Enter the length of time an individual student would be served during one treatment session. A student would be served for thirty minutes in a small group (3 A and B and E = 30), fifty minutes in a large group (3C = 50) and fifteen minutes in an individual tutoring session (3 D and F = 15).

COLUMN 4 - FREQUENCY OF ACTIVITY.

Enter the approximate number of times per week that each seperate activity is served. Take into consideration, if possible, holidays, assemblies, half days and other relevent events that may lessen the frequency. For only the individual instruction sessions account for the average absence rate for the students in addition to the prior considerations.

COLUMN 5 - TARGET STUDENT TREATMENT TIME.

Multiply, column 3, 4 and 5 together for each individual row.

COLUMN 6 - TOTAL NUMBER OF STUDENTS.

Column 6 will generally be the same as column 2 except in those cases were non-title I children are being served with the Title I target students and in case were more than one grade level is being served in the same session as well. In these cases enter the total number of students being served in each seperate activity, regardless of the students level or Title I status.

COLUMN 7 - NUMBER OF INSTRUCTORS.

Enter the number of instructors involved with each activity. If there is more than one instructor in a given treatment session such as a pull out teacher and aide (7B = 2), enter the total number of instructors.

COLUMN 8 - INDIVIDUAL TREATMENT S/T RATIOS.

Divide the number of students (column 6) by the number of instructors (column 7) for each seperate treatment. The large group in row C, as an example, has 21 students (C6 = 21) and one instructor (C7 = 1), thus 21 divided by 1 = 21 (CB = 21). The student/teacher ratio for all individual instruction sessions is automatically one (8 D and F = 1) regardless of how many students are treated.

COLUMN 9 - ADJUSTED STUDENT TREATMENT TIMES.

Divide the Target Student Treatment Time (column 5), by the Individual Treatment S/T Ratios (column 8) for each individual row and enter the figure in column 9. For instance, in row C, Total Student Treatment Time =4200 minutes (C5 =4200), and the Individual Treatment S/T Ratio = 21 (C8 = 21). Thus 4200 divided by 21 = 9.5 (C9 = 240).

COLUMN 10 - TYPE OF INSTRUCTOR.

Simply enter the Title of the instructor or instructors involved in each treatment session.

COMPUTATUON OF THE OVERALL STUDENT / TEACHER RATIO.

To compute the overall student/teacher ratio for any grade follow these steps.

1. Add together all the figures in column 5, (total = 6945)
2. Add together all the figures in column 9, (total = 13311.3)
3. Divide the sum of column 5 by the sum of column 9, to obtain the student teacher ratio = 5.29.

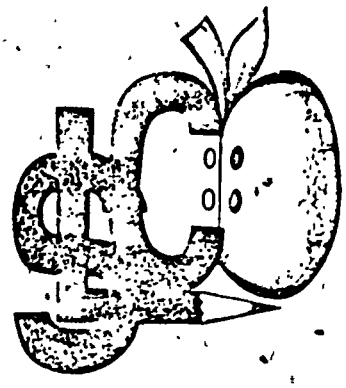
	1 Mode of Activity	2 Number of Students	3 Length of Activity	4 Frequency of Activity	5 Target Student Treatment Time	6 Total Number of Students	7 Number of Instructors	8 Ind. S/T Ratios	9 Adjusted Student Treatment Times	10 Type of Instructor
A	SG	3	30	5	450	7	1	7	64.3	P.0
B	SG	8	30	5	1200	10	2	5	240	P.O. + Aide
C	LG	21	50	4	4200	21	1	21	200	Teacher Leader
D	II	4	15	4	240	11	1	1	240	Aide
E	SG Peer. Tutor	5	30	4	360	5	1	5	72	P.O.
F	II	11	15	3	495	11	1	1	495	P.O.

Student teacher ratio = 6945 ÷ 1311.3 = 5.29

206

Appendix 9

Letters of Approval, Support, & Notification
Regarding Extended Work Scope Project



Salt Lake City School District

440 East First South Salt Lake City, Utah 84111 Phone 322-1471

October 11, 1979

Ms. Cie Taylor
Exceptional Child Center UMC 68
Utah State University
Logan, Utah 84322

Dear Cie,

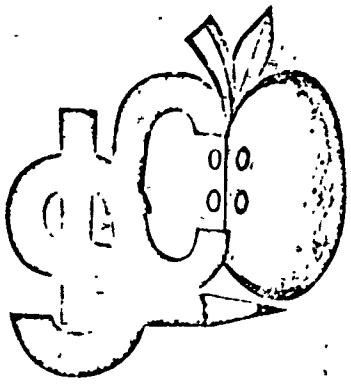
I appreciated the opportunity to talk with you about the problems of group achievement testing with low achieving and learning disabled students. With you, I am concerned that because motivational problems and administration procedures, these test results may not be indicative of the students' true achievement level. Your proposed study sounds like an excellent approach to beginning to provide answers in this important area.

As I explained to you, our District has a committee which must approve all outside research. Before we could give official approval for you to conduct the project in our District it would have to be cleared by this committee. However, because the outcomes of the project would be central to many of our District concerns, I do not anticipate any problems in obtaining this approval.

Good luck with your project! I look forward to hearing more about it from you.

Sincerely,

Darlene Ball
Administrator for Educational Accountability



Salt Lake City School District

440 East First South Salt Lake City, Utah 84111 Phone: 322-1471

April 15, 1980

Mr. Joseph A. Gappa
Office of the Vice President for Research
UMC 14
Utah State University
Logan, Utah 84322

Dear Mr. Gappa:

District personnel in the Salt Lake City School District have reviewed the components of the research proposal entitled The Effect of Reinforcement and Training on Group Standardized Testing Behavior of Mildly Handicapped and High Risk Students.

We grant approval for the implementation of the research as proposed and endorse the efforts of project personnel to increase the validity of group administered standardized instruments.

I have read the Informed Consent Format and understand the elements therein. We feel that the rights and welfare of the subjects (second grade students) will not be violated under the research provisions. To insure parental approval, each parent will be provided with a letter explaining the study and will have an opportunity to withdraw their child from the observation, training, and reinforcement procedure.

Sincerely,

Stanley R. Morgan

Stanley R. Morgan, Administrator
Research and Public Information

SRM:ab

cc: Cie Taylor

UTAH STATE UNIVERSITY · LOGAN, UTAH 84322

801-750-1981

UNIVERSITY AFFILIATED
EXCEPTIONAL CHILD CENTER
UMC 68

April 4, 1980

Dr. Stanley Morgan
440 E 1st S
Board of Education
Salt Lake City School District
Salt Lake City, UT 84111

Dear Dr. Morgan:

Thank you for reviewing and approving the research proposed in The Effect of Reinforcement and Training or Group Standardized Testing Behavior of Mildly Handicapped and High Risk Students. In response to our phone conversation on Monday, March 17, I have enclosed the documents necessary for your affirmation that the rights and welfare of children will not be violated by implementing the project. The following items are included:

1. A copy of the Informed Consent Format. This form is provided for your information and need not be filled in or returned. It is referred to in the letter to be sent to Mr. Gappa.

2. A draft of the letter which is to be sent to Utah State University to assure that the rights and welfare of the students have been protected.

3. A draft of the letter to be sent to each second grade parent (from the principal) to explain the research.

Please read the draft of item 2 above and feel free to edit it to suit your needs. The letter should be sent to Mr. Gappa with a copy to me.

Thank you, Dr. Morgan, for being so helpful. You were so pleasant on the phone, and I hope we can meet in the near future. Please call me collect (750-2044) if I can assist you in any way.

Regards,

Cie Taylor

CT:mmt

Enclosures

April 4, 1980

Dear Parents:

The Salt Lake City Schools are working in cooperation with Utah State University this year on a project designed to investigate the validity of using group standardized tests to measure student academic achievement. The project will train children in test taking strategies and train their teachers in test administration practices. Both of these training programs can eliminate many testing problems by preparing the students and teacher for testing. The tasks of taking and giving tests may also become less difficult and less negative.

A total of 24 second grade classrooms in 12 schools located in Salt Lake City are included in this study. Your child's classroom is one of those participating in this project. Children who participate in the project will be observed during the regularly scheduled District-wide spring testing (April 28 - May 1, 1980). As normal, all test scores will be kept confidential and only group scores will be used to report data.

In order to compare the performance of those students who are trained with those who are not trained, only some students will receive the instruction in test taking. Should your child be chosen for training, one or two hours of instruction is being provided during school hours up to two weeks before the actual testing but will not otherwise interfere with your child's regular work.

In addition to providing some students with training in test taking, all students will have the opportunity to earn a monetary reward for doing well on the test. The average amount of the reward to be given to students will be $1.00. Each child will have an individual goal of a specific test score that is set before the test. If your child attains this individual goal, he or she will earn a reward on the day following the test. Depending on the group to which your child is assigned, these rewards may be earned for math or for reading gains.

Our preliminary results indicate that these training programs will benefit most elementary teachers and pupils in the Salt Lake City Schools. However, if you have any questions regarding this project or the training, please contact the me for further information. If for some reason you would prefer that your child not participate in the study, you may notify the school office and your child will not be included in the training, observation, or reinforcement. Thank you for your cooperation in this project.

Sincerely,

Cie Taylor
Research Assistant

March 31, 1980

Dear

Our meeting on Thursday, March 27, was not only informative but delightful. I am pleased that we had the opportunity to chat for a brief time while reviewing the plans for the spring testing (SAT). I spoke with Maurine McDonald after seeing you and she will be mailing you a testing schedule this week.

In reference to the letters to be sent by you to parents explaining the testing program, you suggested that your school send copies home with the students. I had agreed to provide you with a draft of this letter for your editing. However, through an error in communication the draft was duplicated at the District office and delivered to you in bulk. If you haven't already received this package, it will probably arrive this week.

I apologize for this error. Feel free to change the letter to suit your style and make your own duplications. This letter should be sent home with all second grade students in the classrooms that were chosen to participate in the study. This information is included on the project outline I left with you during my visit.

Thank you for providing so much cooperation in our spring testing project. I will be contacting you shortly regarding the exact scheduling of project activities. Please call me collect (750-2044) if additional concerns arise.

Regards,

Cie Taylor
Research Assistant

CT:dg

Appendix 10

Approval Forms and Letters Related to Videotaping

UTAH STATE UNIVERSITY · LOGAN, UTAH 84322

801-750-1981

UNIVERSITY AFFILIATED
EXCEPTIONAL CHILD CENTER
UMC 68

May 19, 1980

Dear Parent:

Utah State University and Ellis Elementary School are jointly cooperating in a project to produce a short filmed teaching sequence that is designed to improve the test taking skills of students. We have selected your child to participate as a student in the filming. The filming will take place on May 21, 1980 in the second grade classroom during the afternoon session.

Attached is a release form granting permission to film your child and use the videotape for educational purposes. Should you have further questions, please contact your child's teacher.

Yours truly,

Larry Jacobsen

Larry Jacobsen
Principal, Ellis Elementary School

UTAH STATE UNIVERSITY · LOGAN, UTAH 84322

801-750-1981

UNIVERSITY AFFILIATED
EXCEPTIONAL CHILD CENTER
UMC 68

LOCATION: _____ TEACHER: _____

RELEASE TO USE VIDEOTAPES FOR EDUCATIONAL PURPOSES:

I hereby grant permission and authorize Utah State University to take, use and distribute videotapes of me, named below, for the purposes of producing educational information and instructional materials in a manner to be selected by the University. I understand that this includes the right to use and license the use of such videotapes for any educational purpose, including teacher and aide training workshops and observer training sessions. I agree that I will not institute or support any claim or suit of any nature against Utah State University or the persons to whom it might license use or distribution of such pictures.

_____ _____
Date Legal Signature

_____
Address

_____
Phone

March 18, 1980

Dear Parents,

Thanks everyone for being cooperative about the videotape we are making. Some of you have requested an opportunity to see this film. We are tentatively scheduling a viewing for Friday, March 28 at 2:30 P.M. If you are interested and could come this day, please return the bottom of this note.

Thank you again,

Mrs. Fifield.

P.S. They sure do a cute job!

☐ Yes, I'm coming!

☐ I want to come, but have a conflict. A better time would be _____

☐ I don't care to come.

Signature _____